A Time-frequency Masking Based Random Finite Set Particle Filtering Method for Multiple Acoustic Source Detection and Tracking

A Time-frequency Masking Based Random Finite Set Particle Filtering Method for Multiple Acoustic Source Detection and Tracking

Xionghu Zhong, *Member, IEEE*, and James R. Hopgood, *Member, IEEE*

Abstract

Considering that multiple talkers may appear simultaneously, a time-frequency (TF) masking based random finite set (RFS) particle filtering (PF) method is developed for multiple acoustic source detection and tracking. The time-delay of arrival (TDOA) measurements of multiple sources are extracted by using a time-frequency masking technique, by which each source's TF bins are clustered and separated in a joint gain-ratio and time-delay histogram. Since a joint detection and tracking problem is considered, both source positions and source numbers are time-varying and need to be estimated. The tracker is built within a RFS Bayesian filtering framework. Essentially, an RFS process is used to characterize the source dynamics that include source appearance/dissappearance and motion trajectories. Latent variables are also introduced to indicate source dynamics and measurement-source associations. Subsequently, a Rao-Blackwellization PF technique is employed so that the source position state can be marginalized and only the latent variables are estimated by using the PF. The main advantage of the proposed approach is that hypothesis-pruning is formulated in a full probabilblackense. The performance of the proposed approach is demonstrated in real speech recordings as well as in simulated room environments.

Index Terms

Acoustic source tracking, room reverberation, time-delay of arrival, particle filtering, random finite set.

I. Introduction

Acoustic source localization and tracking in a room environment plays an important role in many speech and audio applications such as multimedia, hearing aids, hands-free speech communication, and teleconferencing systems. Once the source or talker is localized and tracked, the position information can be fed into a higher processing stage

Xionghu Zhong is with the School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University, Singapore. 639798. Email: xhzhong@ntu.edu.sg.

James R. Hopgood is with the Institute for Digital Communications, Joint Research Institute for Signal and Image Processing, School of Engineering, The University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK. Email: James.Hopgood@ed.ac.uk

for high-quality speech acquisition, enhancement of a specific speech signal in the presence of other competing talkers, or keeping a camera focused on the talker in a video-conferencing scenario [1]–[6]. Usually, a distributed system equipped with a number of microphone pairs or arrays is employed to localize or track the source [7]–[12]. However, it is difficult to provide an accurate position estimate since the received audio signal can be significantly distorted and its statistical properties can be changed drastically due to room reverberation. The difficulty is further increased when multiple simultaneously active speakers are considered in the tracking scene.

In the past, time-delay of arrival (TDOA) measurements were extensively employed and studied for room acoustic source localization or tracking [8]–[10], [13]–[19] due to their simplicity and ease of access in many applications. TDOA measurements can be extracted, for example, by employing the generalized cross-correlation (GCC) function [20] or adaptive eigenvalue decomposition (AED) algorithm [21]. Since each TDOA yields half a hyperholoid of two sheets which, in the far field, can be approximated by an angular segment, multiple TDOA measurements from distributed microphone arrays are usually employed to triangulate a target position [22], [23]. Such a triangulation can be approximated by either using a linear intersection (LI) algorithm [7] or an extended Kalman filter (EKF) [9], [17]. However, in the presence of noise and room reverberation, ghost peaks may present in the GCC function and spurious TDOA measurements may be collected and the subsequent triangulation methods can be seriously degraded. In [10], [23], [23]–[25], a reverberant measurement model which consists of the TDOAs measurements from real detections as well as false alarms was introduced and the sequential importance resampling based particle filter (SIR-PF) approaches were employed to track the source. In essence, a bi-model hypothesis likelihood was employed to reduce the estimation error due to false TDOA measurements: a Gaussian distribution for real TDOA measurement hypothesis and a uniform distribution for false alarm hypothesis. Generally, the probability of detection can be enhanced due to incorporating multiple TDOA measurement, and the SIR-PF is more robust than the EKF approach in noisy and reverberant environments.

In a real conversation, multiple talkers can be simultaneously active and, under such a scenario, the received signal is a mixture of different speech sounds. This significantly increases the complexity of the tracking problem since: *i*) TDOAs for different sources are no longer easily available; and *ii*) given the TDOA measurements for multiple sources, the measurement-source assignment is unknown. In this paper, a novel approach is proposed to jointly detect and track an unknown and time-varying number of acoustic sources in the room environment. Knowing that traditional GCC methods may not yield sharp peaks for TDOAs of multiple sources, a degenerate unmixing estimation technique (DUET) [26], [27] based GCC (DUET-GCC) method is proposed for TDOA estimation. DUET assumes that the source signals are window-disjoint-orthogonal (WDO) on the time-frequency (TF) domain. Hence, the source's TF spectrum can be separated by clustering all TF bins in a two dimensional (2-D) histogram over time-delay (TD) and gain-ratio (GR) axes. The GCC method is then applied to each source spectrum and the TDOA for each source is obtained. The authors in [28] developed a WDO based KF to localize the DOAs of multiple sources. Although the TDOA measurement extraction approach is similar to the DUET-GCC method developed in this paper, it did not take the phase ambiguity problem into account and can only be applied for arrays with small microphone separations (maximum 4cm in [28]). Mandel et al [29] also built a probabilistic models for

TD and GR information and used an expectation-maximization (EM) algorithm to find the TDOAs of multiple sources. However, the EM algorithm needs a burn-in period to converge to the final estimates, and is thus more appropriate for the localization problem. Other multi-source TDOA estimation methods based on signal separation for localization problem can also be found in [30]–[36].

In the past, various multisensor multitarget tracking techniques were introduced to the multiple talker tracking problem. In [37], [38], direction of arrivals (DOAs) from distributed microphone arrays were obtained and an interacting multiple model (IMM) based probabilistic data association (PDA) technique was developed to fuse the DOAs and estimate the source positions. These approaches assume a perfect detection of the sources and cannot be applied to scenarios where the number of sources are changing and unknown. In [39]–[42], a random finite set (RFS) based Bayesian filtering approach was presented, by which jointly detecting and tracking an unknown and time-varying number of sources is possible. However, these approaches simply exploit the GCC method for TDOA measurement extraction while the GCC method assumes a single source wavefront impinging on a microphone array [20]. More recently [43], an independent component analysis (ICA) based approach was introduced to demix the speech mixtures from multiple sources and a probability hypothesis density (PHD) filter was employed to track the direction of arrivals (DOAs) of the sources. However, such an approach concentrates on DOA estimation using a single microphone array and cannot localize the source exactly in a real room environment. In addition, all these multi-source tracking approaches are neither tested in a broad range of noisy and reverberant environments, nor in a real room experiment. Nonparametric Bayesian approaches were introduced to multiple source localization and separation in [44], [45]. However, the sources considered therein are assumed to be static. In [46], binaural cues (interaural time and intensity differences) were extracted from a microphone pair, and these observations are compared with predicted reference values obtained from simulations using prior knowledge of a catalogue head-related transfer functions (HRTFs). These reference values are obtained based on the binaural response of a KEMAR dummy head. Moreover, a HMM framework is proposed to model the change in the number of active speakers probabilistically. The target space is modeled as a set of subspaces and switches among them with predefined jump probabilities. Again, the method focuses on the DoA tracking rather than tracking of the exact Cartesian $(x - y)$ positions of the sources.

Since the number of sources as well as their respective positions is unknown and time-varying, the source dynamics include source birth (new source activated), source death (existing source nonactivated), and motion of survival sources. In this paper, a RFS particle filter (RFS-PF) [47], [48] is used to detect and track multiple acoustic sources based on TDOA measurements. Following the idea in [47], latent variables are incorporated to identify the motion models and the associations between the TDOAs and the sources. The measurement function is linearized to form an EKF, by which the likelihood can be obtained and the source positions can be marginalized out. The birth model is given as a prior probability and the source death is determined by modelling the expected track length or lifetime using a Gamma distribution [49]. Such a death model assigns higher death probability to the source which is unassociated for a longer time.

The PF implementation of the proposed tracking approach is able to determine the hypotheses in a full probabilistic

sense. The performance of the proposed approach is studied using real speech recordings as well as simulated room environments. Whereas the RFS-PF in [47] simply assumes the source death probability as a constant, we model the length of the source track so that the miss detection due to reverberation has been taken care of. Generally, the larger period the source track is not associated with TDOA measurements, the higher the source death probability. The proposed tracking method is implemented under both DUET-GCC and traditional GCC based TDOA measurements. These two implementations are labeled as DUET proposed and GCC proposed respectively.

The core contributions of this paper are: 1) developing a method to improve high-frequency channel assignment in a TF masking based TDOA estimation method; 2) introducing the Gamma distribution to obtain a controllable death rate in the proposed tracking algorithm such that the algorithm is more robust to the missing detection and speech pauses; and 3) comprehensively studying the performance of the environment as well as a range of simulated reverberant environments. These contributions are detailed in Section II-C, Section IV-A and Section V respectively. The advantage of the proposed DUET-GCC method in TDOA estimation is demonstrated by comparing the performance with that of the PHAT-GCC method. The RFS approach in [41] is also implemented and its performance is compared with the performance of the proposed tracking method. The rest of this paper is organized as follows: in Section II, the DUET-GCC based TDOA measurement extraction is introduced; in Section III, the multiple source tracking problem is formulated; the RFS-PF tracking algorithm is proposed in IV; and the performance of the proposed approach is studied under both simulated and real room environments in Section V. Finally, conclusions are drawn and directions for future work are addressed in Section VI. A table of notations is summarized in table I to illustrate the meaning of variables and symbols in the TDOA estimation method and the tracking algorithm.

II. Multiple source TDOA estimation

In this section, a DUET-GCC method is proposed to estimate the TDOAs for multiple sources.The DUET is used to separate the source in the TF domain and the GCC function is then applied to each source's TF bins to obtain the TDOAs. In particular, the method to improve the TDOA estimation in a high frequency region is explained in Section II-C.

A. TDOA estimation via DUET

Assume that L microphone pairs are deployed to receive the speech signals emitted by M_k speakers at a discrete time step k. Let $\Xi = (k, \omega)$ be a TF bin index and $S_{m,\Xi}$ denote the short time Fourier transform (STFT) of the mth source signal. Ignoring the effect of noise and reverberation, the signal model in the TF domain for the ith microphone of the ℓth pair is

$$Z_{\Xi}(\ell, i) = \sum_{m=1}^{M_k} a_{m,k}(\ell, i) e^{-j\omega\tau_{m,k}(\ell,i)} S_{m,\Xi}, \tag{1}$$

where $a_{m,k}(\ell, i) = 1/4\pi r_{m,k}(\ell, i)$ and $\tau_{m,k}(\ell, i)$ represent the attenuation and the time-delay of the mth source signal at the ith microphone of ℓth microphone pair respectively, with $r_{m,k}(\ell, i)$ denoting the corresponding distance.

TABLE I

NOTATIONS

Algorithm	DUET-GCC
input	$L \times 2$ received microphone signals
output	TDOAs $\cup_{\ell=1}^{L}\{\hat{\tau}_{1,k}(\ell), n = 1, \ldots, n_k(\ell)\}$
hyper par.	A, D
$\Xi = (k, \omega)$	TF bin index at the kth time step and frequency ω
ℓ, L	microphone pair index ℓ; totally L pairs
m,M_k	source index m; totally M_k sources
n, $n_k(\ell)$	measurement index n; totally $n_k(\ell)$ measurements
A, D	GR and TD resolution parameters
$\hat{a}_{n,k}\ell$, $\hat{\tau}_{n,k}(\ell)$	amplitude estimate and TDOA estimate
$\mathcal{R}_{n,k}(\ell, \tau)$	GCC function due to the nth 2-D histogram peak
Algorithm	proposed RFS-PF tracking algorithm
input	$\mathcal{Z}_k = \cup_{\ell=1}^{L}\{\hat{\tau}_{1,k}(\ell), n = 1, \ldots, n_k(\ell)\}$
output	estimated source state set $\hat{\mathcal{X}}_k$
hyper par.	(v, ρ), (α, β), σ_τ
$\mathbf{X}_{m,k}$, $\mathcal{X}_k$	the mth state vector $\mathbf{X}_{m,k}$ and the state set $\mathcal{X}_k$
$z_{n,k}$, $\mathcal{Z}_k$	single measurement $z_{n,k}$ and the TDOA set $\mathcal{Z}_k$
$\mathcal{Z}_{1:k}$	measurements from the start to the time step k
$z_{1:n,k}$	measurements from 1 to n at time step k
$\gamma_{n,k}$	TDOA assignment indicator
(i)	particle index, $i = 1, \ldots, N$
b_k, d_k	birth and death indicators
$\boldsymbol{\theta}_k$	latent variable including the assignment, birth and death indicators
(v, ρ)	Langevin motion model parameters
(α, β)	Gamma distribution parameters
σ_τ	measurement (TDOA) noise variance

According to the WDO assumption [26], the TF bins are disjoint. Hence, each TF bin either carries information regarding one of the sources, or has no meaningful information. The ratio of the TF bins across a microphone pair can be defined as

$$R_\Xi(\ell) = \frac{Z_\Xi(\ell, 1)}{Z_\Xi(\ell, 2)} = a_\Xi(\ell) e^{-j\omega\tau_\Xi(\ell)}, \tag{2}$$

where $a_\Xi(\ell)$ and $\tau_\Xi(\ell)$ are the gain-ratio (GR) and time-delay (TD) estimates for TF bin Ξ respectively. Suppose that the mth source is active on Ξ (the contribution of other sources on this TF bin is thus nil), the GR and TD are given respectively as

$$\begin{aligned} a_\Xi(\ell) &= |R_\Xi(\ell)| = \frac{a_{m,k}(\ell, 1)}{a_{m,k}(\ell, 2)} \triangleq a_{m,k}(\ell), \\ \tau_\Xi(\ell) &= \frac{\angle R_\Xi(\ell)}{-\omega} = \tau_{m,k}(\ell, 1) - \tau_{m,k}(\ell, 2) \triangleq \tau_{m,k}(\ell), \end{aligned} \tag{3}$$

with $|\cdot|$ and $\angle\cdot$ denoting the amplitude and the phase of the estimates respectively, and $a_{m,k}(\ell)$ and $\tau_{m,k}(\ell)$ are the GR and TD information of the mth source separately. Note that the TF bin index Ξ can be omitted in (3) as the GRs and TDs for each source's TF bins are the same. Next, the histogram of all TF bins will be generated based on the GR and TD information and each source's TF bins can thus be clustered and separated in the TF domain.

Given a GR resolution parameter A and a TD resolution parameter D, define an indicator function such that

$$\Lambda_{\Xi}(\zeta,\eta,\ell)=\begin{cases}1, & |a_{\Xi}(\ell)-\zeta A|\leq A \text{ and } |\tau_{\Xi}(\ell)-\eta D|\leq D;\\ 0, & \text{otherwise,}\end{cases}$$

where ζ and η are any integers which lead $(\zeta A,\eta D)$ to cover a GR and TD range completely. The function Λ indicates whether the GR and TD of a TF bin are within the neighborhood of the given parameter pair $(\zeta A,\eta D)$. Based on this indicator, a 2-D histogram for different integers ζ and η can be constructed as

$$h_k(\zeta,\eta,\ell)=\sum_{\Xi}\Lambda_{\Xi}(\zeta,\eta,\ell)\,|Z_{\Xi}(\ell,1)Z_{\Xi}(\ell,2)|^{\gamma}\,, \tag{4}$$

where $|Z_{\Xi}(\ell,1)Z_{\Xi}(\ell,2)|^{\gamma}$ is a weighting term for some γ. For a source with GR and TD parameters $(\zeta A,\eta D)$, a large portion of the source TF bins will carry such GR and TD information and can be clustered in the 2-D histogram. Detailed discussion about the different choices of γ can be found in [26]. Here $\gamma=0$ is picked to equalize the importance of all TF bins, by which the effect of signal energy is reduced.

Assume that $n_k(\ell)$ peaks $(\bar{a}_{n,k}(\ell),\bar{\tau}_{n,k}(\ell))$, for $n=1,\ldots,n_k(\ell)$ above a given threshold can be detected from the 2-D histogram. The indicator function of the TF bins corresponding to the nth peak can be given as

$$\mathbb{I}^{n}_{\Xi}(\ell)=\begin{cases}1, & |a_{\Xi}(\ell)-\bar{a}_{n,k}(\ell)|<A,\ \ |\tau_{\Xi}(\ell)-\bar{\tau}_{n,k}(\ell)|<D;\\ 0, & \text{otherwise,}\end{cases} \tag{5}$$

The TDOA for the nth peak in the 2-D histogram, $h_k(\zeta,\eta,\ell)$, can thus be estimated as

$$\hat{\tau}_{n,k}(\ell)=\mathbb{E}_{\Xi}\Big(\mathbb{I}^{n}_{\Xi}(\ell)\tau_{\Xi}(\ell)\Big), \tag{6}$$

where $\mathbb{E}_{\Xi}(\cdot)$ denotes the expectation over Ξ. The DUET based TDOA measurement extraction is a speech separation based approach. Unlike the traditional GCC function which is unable to differentiate source signals, the TDOAs obtained by DUET are estimated from the TF bins of each source signal. It is thus able to give the TDOAs for multiple sources, even when these sources are simultaneously active. However, in the presence of noise and reverberation, the spectrogram is smeared and blurred. The WDO assumption is thus violated and the TDOA estimation will be degraded. In particular, the expectation step in equation (6) is very sensitive to the TD and GR parameters (A,D). All these factors can make the TDOA estimation diverge from the ground truth, i.e., $\hat{\tau}_{n,k}(\ell)\neq\tau_{m,k}(\ell)$, $\forall m=1,\ldots,M_k$. In the next section, a DUET based GCC method will be developed to obtain robust TDOA estimation.

B. DUET-GCC method

Since the GCC method is found robust in noisy and reverberant environments [20], [50], it is expected that a combination of GCC and the TF bins of each source extracted by DUET to provide reliable TDOA estimates

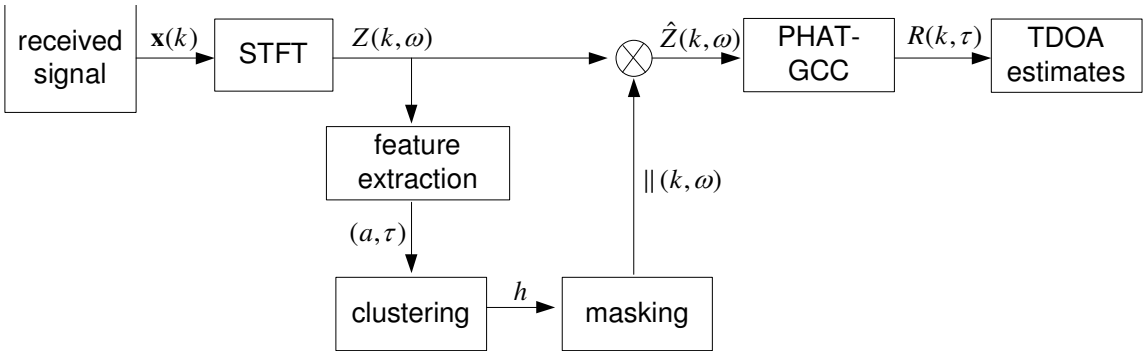


Fig. 1. Flow diagram of the DUET-GCC approach.

for multiple sources. Suppose that $n_k(\ell)$ peaks can be enumerated from a 2-D histogram, and each peak has the indicator $\mathbb{I}_{\Xi}^{n}(\ell)$ defined by (5). The TF bins corresponding to the nth peak, $n = 1, \ldots, n_k(\ell)$ can be given by

$$\widehat{Z}_{\Xi}^{n}(\ell, i) = \mathbb{I}_{\Xi}^{n}(\ell) Z_{\Xi}(\ell, i). \tag{7}$$

The GCC function for the TF bin set can thus be written as

$$\mathcal{R}_{n,k}(\ell, \tau) = \int_{\Omega} \widehat{\Phi}_{\Xi}^{n}(\ell) \widehat{Z}_{\Xi}^{n}(\ell, 1) \widehat{Z}_{\Xi}^{n*}(\ell, 2) e^{j\omega\tau} d\omega, \tag{8}$$

where Ω is the frequency range over which the integral is implemented. In this paper, we have limited the frequency range over $[300, 3700]$Hz to improve the TDOA estimation. The phase transform (PHAT) weighting term is defined as

$$\widehat{\Phi}_{\Xi}^{n}(\ell) = \left| \widehat{Z}_{\Xi}^{n}(\ell, 1) \widehat{Z}_{\Xi}^{n*}(\ell, 2) \right|^{-1}, \tag{9}$$

where the superscript $*$ denotes the complex conjugate. The TDOA for each source is thus obtained by exploring an one-dimensional search over the TDOA range, given as

$$\hat{\tau}_{n,k}(\ell) = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} \mathcal{R}_{n,k}(\ell, \tau). \tag{10}$$

The TDOA estimation for multiple sources is thus achieved. It is worth pointing out that the DUET-GCC estimation steps (8) to (10) are the same as the traditional GCC estimation procedure, but the spectrogram of each source is employed to replace the spectrogram of whole speech mixtures in the traditional GCC approach.

Figure 1 gives a flow diagram of the DUET-GCC approach. The speech mixtures are separated by using DUET, and the GCC method is then employed for the spectrogram of each source to estimate the TDOAs. Since the TDOA estimation of the speech sources are handled separately, the interference between the source signals is naturally decreased. The DUET-GCC method is thus more appropriate for the TDOA estimation of multiple simultaneously active sources than the traditional GCC method. Also due to the capability of suppressing the reverberation and noise by the PHAT weighting, the TDOA estimation performance via DUET-GCC approach is better than simply taking the expectation of the TDOA information from all the TF bins in equation (6) [25]. Fig. 2 shows the GCC function extracted from DUET-GCC method and traditional PHAT-GCC method respectively. The two largest local

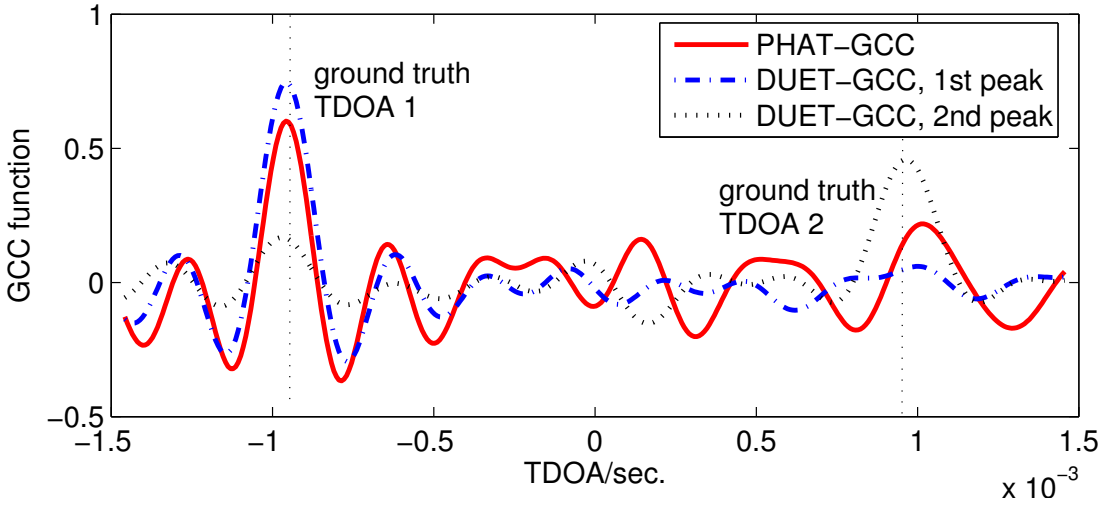


Fig. 2. DUET-GCC function and traditional GCC function via PHAT weighting. Two sources are located at $(1.4, 1.2)$m and $(1.4, 2.8)$m respectively. The GCC function is estimated from the first microphone pair (microphone 1 and microphone 2), as shown, in Fig. 3. The ground truth TDOAs are ± 0.95ms.

peaks in DUET 2-D histogram are used to obtain TF masks, and thus two GCC functions are available. The DUET-GCC method presents accurate TDOA estimates for two sources, while the traditional PHAT-GCC is only able to present one source accurately, and fails to produce a sharp peak for the other one.

C. Practical issues arising from DUET

The microphone separation d by which the phase term will not be wrapped is

$$d < d_{\max} = \frac{c}{2f_{\max}} = \frac{c}{f_{\mathrm{s}}}, \tag{11}$$

where $f_{\max}$ and f_{s} indicate maximum frequency component and sampling frequency of the signal respectively, where c is the sound speed in the air. In many real applications, since the microphone separation will be larger than $d_{\max}$, a phase unwrapping method is needed to unwrap the phase term at the higher frequency band $f > c/(2d_{\max})$. Several approaches have been proposed for this purpose, and a typical one is found in Tribolet's work [51]. Normally the phase estimates are very unlikely to be unwrapped at the low frequency band. Let $\phi_{\Xi}(\ell) = \angle R_{\Xi}(\ell)$ represent the phase of the TF bin ratio at Ξ. First, the phase extracted from the TF bins at the lower frequency band, i.e., $\omega/(2\pi) < c/(2d_{\max})$, are used to build a histogram to obtain initial TDOA estimates $\hat{\tau}_{n,k}(\ell)$ for multiple sources by using DUET. The estimated TDOAs given by equation (6) are then used to predict the phase term at the higher frequency band. The phase at the higher frequency band can be obtained by using linear prediction as $\phi_{\Xi}^{\mathrm{pred}}(\ell) = \omega\hat{\tau}_{n,k}(\ell) + 2\kappa\pi$ for $\omega/(2\pi) > c/(2d_{\max})$. The integer κ is determined by

$$\hat{\kappa} = \arg\min_{\kappa} |\phi_{\Xi}(\ell) - \phi_{\Xi}^{\mathrm{pred}}(\ell) + 2\kappa\pi| \tag{12}$$

for all $\omega/(2\pi) > c/(2d_{\max})$. The unwrapped phase estimates are thus $\hat{\phi}_\Xi(\ell) = \phi_\Xi(\ell) - 2\hat{\kappa}\pi$. After phase unwrapping, the estimated phase $\hat{\phi}_\Xi(\ell)$ at the higher frequency band together with the phase $\phi_\Xi(\ell)$ at the lower frequency band are employed to form the indicator function in equation (5).

In practice, the ideal TD and GR histogram can rarely be achieved due to: i) the outliers of the GR $a_\Xi(\ell)$, which may present in the GR features and make the corresponding TD features not be clustered even though these TD features are correct; and ii) different bins of the TD and GR features, for the TD feature, it is very small compared to the GR feature and thus very difficult to give a meaningful parameter D to cluster them. The detailed studying of TD and GR features can be found in [52]. One way to solve this problem is normalizing these two features [52], and thus make the parameter studying of (A, D) more controllable. Such a normalization is given by

$$\tilde{\tau}_\Xi(\ell) = \frac{\tau_\Xi(\ell)}{2\tau_{\max}}, \quad \text{and,} \quad \tilde{a}_\Xi(\ell) = \frac{|Z_\Xi(\ell,1)|}{\sqrt{\sum_{i=1}^{2}|Z_\Xi(\ell,i)|^2}}. \tag{13}$$

Hence, the TD feature is within the range $\tilde{\tau}_\Xi(\ell) \in [-1/2, 1/2]$ and GR feature $\tilde{a}_\Xi(\ell) \in (0, 1)$. The TD feature is regarded as a correct estimation of the real TDOA if it is located in the admissible range of anomaly error ϵ. The spacing parameter D is thus picked as a normalized version of anomaly error ϵ, given as

$$D = \frac{\epsilon}{2\tau_{\max}} = \frac{cT_c}{4d_{\text{ref}}}, \tag{14}$$

where d_{ref} is a reference distance that can be chosen as 1m, and T_c is the correlation time defined as the time period that the highest peak of the cross correlation function drops off by 3dB. It has been shown that the cross correlation time is about two samples time interval under a sampling frequency of $f_s = 8$kHz [25]. It is worth mentioning that the TD spacing parameter obtained from (14) is normalized and such a spacing parameter can be applied to microphone pairs with different microphone separations. For a sampling frequency of 8kHz, the normalised TD spacing parameter is thus about 0.02. Since the GR feature is also normalised within a range of one, we can simply set the GR parameter the same as TD parameter, i.e., $A = D = 0.02$. According to our experimental study, slightly changing these parameters will not result in significant variation in the performance of the TDOA estimation.

III. RFS BASED TRACKING FRAMEWORK

The RFS has been shown to be an efficient framework for multiple source tracking since it naturally depicts the randomness of the source number as well as the source positions. In this section, the RFS framework for multiple acoustic source detection and tracking is formulated. In addition, the state to be estimated is decomposed to the source state and the data association variables so that the former can be marginalized out by using the EKF [22] and only the later needs to be estimated by using the PF.

A. RFS framework formulation

Assume that at the ℓth $(\ell = 1, \ldots, L)$ microphone pair, a set of TDOAs $\hat{\boldsymbol{\tau}}_k(\ell) = \{\hat{\tau}_{1,k}(\ell), \ldots, \hat{\tau}_{n_k(\ell),k}(\ell)\}$ is obtained by using the DUET-GCC method at time step k. Such a TDOA set contains the source generated TDOAs

as well as TDOAs due to reverberation and noise. The complete measurement set $\mathcal{Z}_k$ over all microphone pairs can be stated as

$$\mathcal{Z}_k = \bigcup_{\ell=1}^{L} \{\hat{\boldsymbol{\tau}}_k(\ell)\} = \underbrace{\{\hat{\tau}_{1,k}(\ell), \ldots, \hat{\tau}_{n_k^{\mathrm{s}}(\ell),k}(\ell)\}}_{\text{source generated}} \cup \underbrace{\{\hat{\tau}_{1,k}(\ell), \ldots, \hat{\tau}_{n_k^{\mathrm{f}}(\ell),k}(\ell)\}}_{\text{false alarms}}, \tag{15}$$

where $n_k^{\mathrm{s}}(\ell)$ and $n_k^{\mathrm{f}}(\ell)$ represent the number of source generated measurements and the number of false alarms respectively. The cardinality of the measurement set is thus $N_k = |\mathcal{Z}_k| = \sum_\ell n_k(\ell) = n_k^{\mathrm{s}}(\ell) + n_k^{\mathrm{f}}(\ell)$ with $|\cdot|$ standing for the cardinality. Since the data association is considered, the measurement set from all microphone pairs are processed sequentially. The measurement set for each step is thus a singleton, given as $z_{n,k} = \{\hat{\tau}_{n,k}\}$, for all $n = 1, \ldots, N_k$. Here, the microphone pair index ℓ is dropped as the TDOA measurements are processed in an arbitrary order. In the following, the expression $\mathcal{Z}_{1:k}$ will refer to all the measurements from the start to the current time step, and $z_{1:n,k}$ corresponds to the measurements from 1 to n at time step k.

For joint detection and tracking problem, other than modeling the trajectory of each source, more complicated dynamic models should be incorporated to take the uncertainty of the source appearance and disappearance into account. Three categories of the source behaviors are considered for the speaker detection and tracking problem: source survival, source birth and source death. Any kind of source dynamics can thus be modelled by formulating a combination of these three behaviors. The motion of the sources can assumed to be independent. Hence, the Langevin model [10] which has been shown to be effective in modeling simple trajectory and slow-paced movements, is employed to model the trajectory of each source. Let $\mathbf{x}_{m,k} = [x_{m,k}, y_{m,k}]^T$ and $\dot{\mathbf{x}}_{m,k} = [\dot{x}_{m,k}, \dot{y}_{m,k}]^T$ denote the position and velocity of the mth source respectively. The superscript T represents the transpose. The complete state characterizing the motion of the source is thus $\mathbf{X}_{m,k} = [\mathbf{x}_{m,k}^T, \dot{\mathbf{x}}_{m,k}^T]^T$. Let $\mathcal{X}_{k-1} = \{\mathbf{X}_{1,k-1}, \ldots, \mathbf{X}_{M_{k-1},k-1}\}$ represent the state set at time step $k-1$. If there is no birth and death, the state set is $\mathcal{X}_{k|k-1} = \bigcup_{m=1}^{M_{k-1}} \{\mathbf{X}_{m,k|k-1}\}$, in which $\mathbf{X}_{m,k|k-1}$ is evolved from the previous state $\mathbf{X}_{m,k-1}$ following the Langevin model, given by

$$\mathbf{X}_{m,k|k-1} = \mathbf{A}\mathbf{X}_{m,k-1} + \mathbf{Q}\mathbf{v}_k. \tag{16}$$

The coefficient matrices $\mathbf{A}$ and $\mathbf{Q}$ are given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_2 & a\Delta T \mathbf{I}_2 \\ \mathbf{0} & a\mathbf{I}_2 \end{bmatrix}; \quad \mathbf{Q} = \begin{bmatrix} b\Delta T \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & b\mathbf{I}_2 \end{bmatrix}, \tag{17}$$

where $\Delta T = T_0/f_s$ is the time interval (in seconds) between time step k and $k-1$, f_s denoting the sampling frequency, and $\mathbf{I}_M$ is an M-order identity matrix. The parameters a and b are the position and velocity variance constants calculated according to $a = \exp(-\rho\Delta T)$ and $b = v\sqrt{1-a^2}$, in which v and ρ are the velocity parameter and the rate constant respectively. The model parameters $v = 1\mathrm{ms}^{-1}$ and $\rho = 10\mathrm{s}^{-1}$ used in [10], [24], [41] are found to be adequate for room acoustic source tracking are employed here. Since there is no source birth or death, the number of the sources remains M_{k-1}, i.e., $|\mathcal{X}_{k|k-1}| = |\mathcal{X}_{k-1}|$.

The birth process happens when a new speaker becomes active in the tracking scene. Let $\mathcal{B}_k$ be the set of new born sources. To simplify the problem, only one source is allowed to be born at each time step. Assume that the new source is born with an initial state $\mathbf{X}_0$. The birth process is then $\mathcal{B}_k = \{\mathbf{X}_0\}$. The total number of the sources at time step k is thus $M_k = M_{k-1} + 1$. When an existing source becomes silent, the death process is formulated by removing the corresponding state from the existing set. As with the birth process, we assume that maximum one source is allowed to die at each time step.

B. Tracking via data association

Assume that there are M_k sources at time step k, i.e. $|\mathcal{X}_k| = M_k$. For each singleton measurement set, $z_{n,k}$, an assignment hypothesis variable $\gamma_{n,k}$ is defined to identify the association between the measurement and the source: $\gamma_{n,k} = 0$ denotes that $z_{n,k}$ is a false alarm; $\gamma_{n,k} = m$, for $m = 1, \ldots, M_k$ denotes that $z_{n,k}$ is associated to the mth source; and $\gamma_{n,k} = M_k + 1$ denotes that $z_{n,k}$ is associated to a new born source labeled as $M_k + 1$. At each time step, the assignment variable set is thus $\boldsymbol{\gamma}_k = \{\gamma_{1,k}, \ldots, \gamma_{N_k,k}\}$. Each measurement can only be associated with one source. Suppose that $\boldsymbol{b}_k$ and $\boldsymbol{d}_k$ are the variables indicating the birth and the death processes of the source respectively, with value 1 denoting that the birth and death happen and 0 otherwise. Defining a latent variable $\boldsymbol{\theta}_k = (\boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k)$, the complete state set for source dynamics is given as

$$\mathcal{Y}_k = \mathcal{X}_{k|k-1} \bigcup \{\boldsymbol{\theta}_k\}. \tag{18}$$

where $\mathcal{X}_{k|k-1}$ is the predicted state given in Section III-A.

The aim here is to estimate the joint posterior distribution $p(\mathcal{Y}_k|\mathcal{Y}_{k-1}, \mathcal{Z}_{1:k})$, which can be decomposed into the conditional source distribution $p(\mathcal{X}_k|\mathcal{X}_{k-1}, \boldsymbol{\theta}_k, \mathcal{Z}_{1:k})$ and the association posterior density $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}, \mathcal{Z}_{1:k})$, given by

$$p(\mathcal{Y}_k|\mathcal{Y}_{k-1}, \mathcal{Z}_{1:k}) = \underbrace{p(\mathcal{X}_k|\mathcal{X}_{k-1}, \boldsymbol{\theta}_k, \mathcal{Z}_{1:k})}_{\text{EKF approximation}} \underbrace{p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}, \mathcal{Z}_{1:k})}_{\text{PF}}. \tag{19}$$

Conditional on $\boldsymbol{\theta}_k$, the position states $p(\mathcal{X}_k|\mathcal{X}_{k-1}, \boldsymbol{\theta}_k, \mathcal{Z}_{1:k})$ can be estimated using an EKF [22], and only the latent variable $\boldsymbol{\theta}_k$ is needed to be handled by a PF. Such a technique is referred as a Rao-Blackwellization PF (RBPF) and widely used for the state estimation where part of state can be marginalized out analytically [53]. The idea of the RBPF is marginalizing out some of the variables by using an optimal filter and estimating the rest by using a particle approximation so that the dimension of the state to be sampled can be reduced. Consequently, using a smaller number of particles is able to achieve the same accuracy of the state estimation as the PF [54]. Suppose that N particles $\boldsymbol{\theta}_k^{(i)}$, for $i = 1, \ldots, N$ are drawn according to the importance distribution of the latent variable $\boldsymbol{\theta}_k$

$$\boldsymbol{\theta}_k^{(i)} \sim q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k}). \tag{20}$$

where $q(\cdot)$ is the importance function that will be detailed in Section IV.B. Under a RBPF implementation, the weight of the particles can be updated as [55]

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k|\boldsymbol{\theta}_k^{(i)}, \mathcal{Z}_{1:k-1}) p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})}{q(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k})}, \tag{21}$$

Algorithm 1: Top-level procedure of the RBPF.

Initialization: $w_0^{(1:N)} \leftarrow 1/N$; $\mathcal{X}_0^{(1:N)} \leftarrow \emptyset$.

for $k \leftarrow 1$ **to** K **do**

1) predict the state $\hat{\mathcal{X}}_{k|k-1}^{1:N}$according to (16).

for $n \leftarrow 1$ **to** N_k **do**

for $i \leftarrow 1$ **to** N **do**

2) generate different hypothesis $\boldsymbol{\theta}_k^{(i)}$;

3) evaluate the importance function $w_k^{(i)}$ according to (21); see Algorithm 2 for details.

end

4) weight normalization: $w_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N} w_k^{(i)}$.

end

5) Output the estimates.

6) Resample ($\mathcal{X}_k$, w_k) if necessary.

end

where $p(z_k|\boldsymbol{\theta}_k^{(i)}, \mathcal{Z}_{1:k-1})$ is the likelihood of the hypothesis, and $p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})$ is the corresponding prior density. After the resampling step, the association posterior density $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}, \mathcal{Z}_{1:k})$ is approximated by $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}, \mathcal{Z}_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k)$, where $\delta(\cdot)$ is a dirac function only with value one when $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{(i)}$ and 0 otherwise. The joint posterior distribution for the whole state can thus be obtained by

$$p(\mathcal{X}_k, \boldsymbol{\theta}_k|\mathcal{X}_{k-1}, \boldsymbol{\theta}_{k-1}, \mathcal{Z}_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k) p(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)}, \boldsymbol{\theta}_k^{(i)}, \mathcal{Z}_{1:k}), \tag{22}$$

where $\mathcal{X}_k^{(i)}$, for $i = 1, \ldots, N$ are the particle representation of the source states and $p(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)}, \boldsymbol{\theta}_k^{(i)}, \mathcal{Z}_{1:k})$ is the filtered distribution obtained from the EKF step.

The top-level procedure of RBPF is given in Algorithm 1. By using the RBPF, the source states are analytically marginalized by using the EKF, and the dimension of the state to be processed by using the PF is significantly reduced. Hence, the same accuracy can be achieved by using a smaller number of particles. The detailed implementation of each step will be introduced in the next section.

IV. RFS-PF Algorithm for detection and tracking

This section presents the RFS-PF tracking algorithm. The prior densities of the source birth and death processes and the measurement-source association are given first. The likelihood and optimal importance function are then derived. Considering that the TDOA measurements are seriously deteriorated in the heavy reverberant and noisy environment, it is likely that none of the TDOA measurements will be the real detection when the distance between the source and the microphone array is large. Further, speech pauses can also result in miss detection. Under such a

situation, the source actually exists in the tracking scene while the TDOA measurement cannot be detected. To lock onto the source tracks, a Gamma distribution is employed to model the source death prior such that the longer the track is not associated with any TDOA measurements, the larger is the death probability. Such a gamma-distributed lifetime of sound sources is able to flexibly control the death rate and is detailed in Section IV-A.

Several assumptions are made to reduce the exhaustive associations in the variable $\boldsymbol{\theta}_k^{(i)}$: *i*) at most one source can be born at a time step k and the source can only be generated within the boundary of the room; *ii*) at most one source can die at a time step k; and *iii*) the total number of sources is bounded at $N_{\max}$. The restriction of at most one source can be born or die at a time step k is to guarantee that the association and the combinations are always limited. In practice, the number of simultaneously active speakers can be assumed to be small, and thus the maximum number of the sources is bounded to $N_{\max}$ to reduce unnecessary associations. This is easily obtained by set the birth probability as 0 when the maximum number of sources achieves, i.e., $|\mathcal{X}_{k-1}^{(i)}| = N_{\max}$. These assumptions can avoid an exponential increasing of the complexity of the algorithm and thus make the algorithm computationally affordable.

A. Birth, death and association priors

To calculate the prior of the hypothesis variable $p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})$, the relation between the birth, survive, and death process should be clarified. A source is born with a prior birth probability, and is independent with any of the existing sources. Generally, the probability of a source death is dependent only on the previous existence of the source. The measurement to source association is dependent only on the number of sources based on the birth and death assumptions at current time step k. The prior of the association variable can thus be written as

$$\begin{aligned} p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{1:k-1}^{(i)}, \mathcal{Z}_{1:k-1}) =& p(\boldsymbol{b}_k^{(i)}) p(\boldsymbol{d}_k^{(i)}|\boldsymbol{d}_{k-1}^{(i)}) \\ & \times p(\boldsymbol{\gamma}_k^{(i)}|\boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)}), \end{aligned} \tag{23}$$

where $p(\boldsymbol{b}_k^{(i)})$ and $p(\boldsymbol{d}_k^{(i)}|\boldsymbol{d}_{k-1}^{(i)})$ are the prior density of the birth and death processes respectively, and $p(\boldsymbol{\gamma}_k^{(i)}|\boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)})$ is the prior density of the measurement-source associations. Detailed expressions of these priors are given as follows.

The birth process happens with predefined probability of P_{b} and independent with any existing sources. The probability of a birth process can thus be given as

$$p(\boldsymbol{b}_k^{(i)}) = \begin{cases} P_{\mathrm{b}}, & \boldsymbol{b}_k^{(i)} = 1; \\ 1 - P_{\mathrm{b}}, & \boldsymbol{b}_k^{(i)} = 0 \text{ or } |\mathcal{X}_{k-1}^{(i)}| = N_{\max}. \end{cases} \tag{24}$$

In practice, P_{b} is unknown and is usually determined by experimental study. Generally, increasing the value of the birth probability P_{b} is expected to enhance the probability of discovering a new source. However, an overly large value may increase the risk of overestimation of the source number.

Given an existing source in the tracking scene, its lifetime is modelled to obtain a death prior. In this work, the lifetime of a source T_m is modeled by a gamma distribution [49]. The gamma probability density function is widely used in reliability models of lifetimes, and is more flexible than the exponential distribution in that it can

be regarded as a summation of multiple exponential distributions and can be used to model the variables that seem to be highly skewed. The probability of the expected track length of the mth source follows a gamma distribution, given as

$$T_m \sim \mathcal{G}(T_m|\alpha,\beta) = {T_m}^{(\alpha-1)}\frac{\beta^\alpha e^{\frac{-T_m}{\beta}}}{\Gamma(\alpha)}, \tag{25}$$

where $\mathcal{G}(\cdot|\alpha,\beta)$ is the gamma distribution with α and β denoting the shape parameter and scale parameter respectively.

Suppose that t_0 is the frame length (in seconds) of the processed speech signal and at time step k, the time stamp is $t_k = kt_0$. Further assume that $t_m^{(i)}$ is the last time that the source m is associated with a TDOA measurement. During the period $\Delta t_m^{(i)} = t_{k-1} - t_m^{(i)}$, the source is not associated but remains in the scene. Given the condition $T_m^{(i)} \geq t_{k-1} - t_m^{(i)}$, we are interested in the probability that the source is dead at current time step t_k, with $t_k = t_{k-1} + t_0$. The probability that the source is dead at current time t_k is [49]

$$p(\boldsymbol{d}_k^{(i)}|\Delta t_m^{(i)}) = P(T_m^{(i)} \in \left[\Delta t_m^{(i)}, \Delta t_m^{(i)} + t_0\right] | T_m^{(i)} \geq \Delta t_m^{(i)}). \tag{26}$$

Given a Gamma distribution, the death probability is determined by the period that it is not associated but still alive $\Delta t_m^{(i)} = t_{k-1} - t_m^{(i)}$. Normally, the larger is $\Delta t_m^{(i)}$, the higher possibility the source m dies. The gamma parameter pair (α,β) controls how fast the source dies.

The source survival are constructed after considering the death process. If the mth source is not dead at current time step k, it is surviving with a probability of $1 - p(\boldsymbol{d}_k^{(i)}|\Delta t_m^{(i)})$. Hence, the density after considering the death process is

$$p(\boldsymbol{d}_k^{(i)}|\boldsymbol{d}_{k-1}^{(i)}) = \sum_{m=1}^{M_k^{(i)}} p(\boldsymbol{d}_k^{(i)}|\Delta t_m^{(i)}) \prod_{\substack{m'=1,\\ m'\neq m}}^{M_{k-1}} \left\{1 - p(\boldsymbol{d}_k^{(i)}|\Delta t_{m'}^{(i)})\right\}. \tag{27}$$

After the birth and death processes, the prior probability of the association indicator can be defined as

$$p(\boldsymbol{\gamma}_k^{(i)} = \gamma|\boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)}) = \begin{cases} p_f, & \gamma = 0; \\ \frac{1-p_f}{M_k^{(i)}}, & \gamma = m, m \neq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

where p_f is the prior probability of false alarm, and the source number $M_k^{(i)}$ is defined as

$$M_k^{(i)} = \begin{cases} |\mathcal{X}_k^{(i)} \cup \mathcal{B}_k| = M_{k-1}^{(i)} + 1, & \boldsymbol{b}_k^{(i)} = 1; \\ |\mathcal{X}_k^{(i)} \setminus \mathcal{D}_k| = M_{k-1}^{(i)} - 1, & \boldsymbol{d}_k^{(i)} = 1; \\ |\mathcal{X}_k^{(i)}| = M_{k-1}^{(i)}, & \text{otherwise.} \end{cases} \tag{29}$$

Since the probability of false alarm is p_f, the probability for all the sources is $1 - p_f$. To keep the summation of the association prior to be unity, a reasonable choice for setting the probability for each source is thus to distribute the probability of all the sources equally, i.e., $(1 - p_f)/M_k^{(i)}$.

B. Optimal importance function

The optimal importance function [56], $q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k}) = p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k})$, has been proved to be able to minimize the variance of the importance weight $w_k^{(i)}$ conditional upon the previous states and measurements. Since the position states $\mathcal{X}_k$ is marginalized out by the EKF, the measurement is only conditional on the latent variable $\boldsymbol{\theta}_k^{(i)}$. The optimal importance distribution can be stated as

$$\begin{aligned}\boldsymbol{\theta}_k^{(i)} &\sim q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k}) \\ &= \frac{p(\mathcal{Z}_k|\boldsymbol{\theta}_k^{(i)}, \mathcal{Z}_{1:k-1})p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})}{p(\mathcal{Z}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})},\end{aligned} \tag{30}$$

Substituting the optimal importance function (30) into the weight updating equation (21), we can get the new expression of the weight updating, given as

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(\mathcal{Z}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1}). \tag{31}$$

where

$$\begin{aligned}p(\mathcal{Z}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1}) &= \int p(\mathcal{Z}_k, \boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1}) d\boldsymbol{\theta}_k \\ &= \sum_{\boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k} p(\mathcal{Z}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k, \mathcal{Z}_{1:k-1}) \\ &\quad \times p(\boldsymbol{\gamma}_k|\boldsymbol{d}_k, \boldsymbol{b}_k, \boldsymbol{\theta}_{k-1}^{(i)}) p(\boldsymbol{d}_k|\boldsymbol{d}_{k-1}^{(i)}) p(\boldsymbol{b}_k),\end{aligned} \tag{32}$$

where $p(\boldsymbol{b}_k)$, $p(\boldsymbol{d}_k|\boldsymbol{d}_{k-1}^{(i)})$, and $p(\boldsymbol{\gamma}_k|\boldsymbol{d}_k, \boldsymbol{b}_k, \boldsymbol{\theta}_{k-1}^{(i)})$ are given in (24), (27) and (28) respectively.

The TDOA measurements are processed one after another. Assume that the measurement at current process is $z_{n,k} \in \mathcal{Z}_k$. Given a latent variable $\boldsymbol{\theta}_k^{(i)}$, the likelihood $p(z_{n,k}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1})$ can be calculated as

$$\begin{aligned}&p(z_{n,k}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1}) = \\ &\qquad \int_{\mathcal{F}} p(z_{n,k}, \mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1}) \mu(d\mathcal{X}),\end{aligned} \tag{33}$$

where the subscript $\mathcal{F}$ is the collection of all finite subsets of the state space, and $\mu(d\mathcal{X})$ is a measure on $\mathcal{F}$. Note that in the case that μ is the Lebesgue measure, $\mu(d\mathcal{X})$ is the same as $d\mathcal{X}$. Since this work is from an application point of view, readers are referred to [57], [58] for a detailed RFS definition and derivation of these PDFs. The above likelihood (33) can be decomposed as

$$\begin{aligned}&p(z_{n,k}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1}) = \\ &\int_{\mathcal{F}} p(z_{n,k}|\gamma_{n,k}^{(i)}, \mathcal{X}) p(\mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1}) \mu(d\mathcal{X}).\end{aligned} \tag{34}$$

If the measurement is associated with a clutter, i.e., $\gamma_k^{(i)} = 0$, $p(z_{n,k}|\gamma_{n,k}^{(i)}, \mathcal{X})$ follows a uniform distribution over the possible TDOA interval is given

$$p(z_{n,k}|\gamma_{n,k}^{(i)} = 0) = \mathcal{U}_{[-\tau_{\max}, \tau_{\max}]}(z_{n,k}) = \frac{1}{2\tau_{\max}}, \tag{35}$$

Algorithm 2: Calculate the optimal importance function.

// over all source states in $\mathcal{X}_{k|k-1}^{(i)}$.

1) calculate the likelihood (38) and the filtered state according to the EKF steps in Appendix A.

2) calculate the likelihood for false alarms from (35).

// over all hypotheses in $\boldsymbol{\theta}_k^{(i)}$.

3) compute the likelihood according to (32).

4) select the hypothesis $\theta_k^{(i)}$, and update the states $\mathcal{X}_k^{(i)}$ accordingly;

5) update the particle weight according to (31).

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay which can only happen when the microphone pair and the source lie exactly on a line. The above expression (34) becomes

$$p(z_{n,k}|\gamma_{n,k}^{(i)} = 0) \int_{\mathcal{F}} p(\mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}, \mathcal{Z}_{1:k-1})\mu(d\mathcal{X}) = \frac{1}{2\tau_{\max}}. \tag{36}$$

In the case that the measurement $z_{n,k}$ is associated with a source, i.e., $\gamma_k^{(i)} = m \geq 0$, for $m = 1, \ldots, M_k$, it follows a nonlinear relationship with the source state given by

$$\hat{\tau}_{m,k}(\ell) = \frac{\|\widehat{\mathbf{x}}_{m,k}^{(i)} - \boldsymbol{p}_{\ell,1}\| - \|\widehat{\mathbf{x}}_{m,k}^{(i)} - \boldsymbol{p}_{\ell,2}\|}{c}. \tag{37}$$

where $\boldsymbol{p}_{\ell,i}$, $i \in \{1, 2\}$ is the position of the ith microphone of the ℓth pair, and $\widehat{\mathbf{x}}_{m,k}^{(i)}$ the state estimate given by (49e) after the EKF step in Appendix A. The likelihood is given by

$$p(z_{n,k}|\gamma_{n,k}^{(i)} = m, \widehat{\mathbf{x}}_{m,k}^{(i)}) = \mathcal{N}(z_{n,k}; \hat{\tau}_{m,k}(\ell), \mathbf{S}_{m,k}^{(i)}), \tag{38}$$

where $\mathbf{S}_{m,k}^{(i)}$ is given by (49c) in Appendix A. The integral in (33) can be written as

$$\begin{aligned} &\int p(z_{n,k}|\gamma_k^{(i)} = m, \mathbf{x}_{m,k}^{(i)}) p(\mathbf{x}_{m,k}^{(i)}|\boldsymbol{\theta}_k^{(i)}, z_{1:n-1,k}) d\mathbf{x}_{m,k}^{(i)} \\ &= p(z_{n,k}|\gamma_{n,k}^{(i)} = m). \end{aligned} \tag{39}$$

Further, if the measurement is associated with a new born source, the same EKF implementation will apply. The formulation of the likelihood follows (38) but using $\mathbf{x}_0$ and $\mathbf{P}_0$ as the state and variance respectively in the EKF implementation.

The calculation of the integration in equation (32) is simply the summation of the probabilities of all the hypotheses. The likelihood can be computed by using equation (35) or equation (38). The EKF also provides the filtered position state distribution if the measurement is associated to a source. The filtered distribution after the EKF implementation is given by equation (50) in Appendix A. The algorithm of the importance function calculation is summarised in Algorithm 2. The advantage of using a PF here is that it allows a random hypothesis pruning/determing rather than heuristic hypothesis selection in traditional gating based data association.

C. State estimation and performance evaluation

The posterior distribution obtained by the RBPF is a joint distribution. Extracting a final state estimation is not as straightforward as that in the single source scenario. The histogram like visualisation of the probability hypothesis density (PHD) can be obtained as [47]

$$D(\widehat{\mathbf{x}}_k) = \sum_{i=1}^{N} w_k^{(i)} \sum_{m=1}^{M_k^{(i)}} \mathcal{N}(\mathbf{x}_k; \widehat{\mathbf{x}}_{m,k}^{(i)}, \widehat{\mathbf{P}}_{m,k}^{(i)}), \tag{40}$$

where $M_k^{(i)}$ is the dimension of the ith particle $\mathcal{X}_k^{(i)}$, and $\widehat{\mathbf{x}}_{m,k}^{(i)}$ and $\widehat{\mathbf{P}}_{m,k}^{(i)}$) are the state vector and the variance matrix of the mth source in the ith particle respectively.

The following error metrics are considered to evaluate the estimation performance over many Monte Carlo (MC) runs: the percentage the tracking algorithm can estimate the right number of sources, and given the correct estimation of the source number, how far the number and position estimates deviate from the ground truth. Suppose that J MC runs are implemented. Let $\widehat{\mathcal{X}}_{j,k}$, $j = 1, \ldots, J$ and $\mathcal{X}_k$ represent the state estimates of the jth run and the ground truth respectively, and $\widehat{M}_{j,k} = |\widehat{\mathcal{X}}_{j,k}|$ is the source number estimation. The probability of the correct number estimation is defined as

$$P_k = \frac{1}{J} \sum_{j=1}^{J} \delta_{|\mathcal{X}_k|} \left(|\widehat{\mathcal{X}}_{j,k}| \right) \times 100\%. \tag{41}$$

The probability of correct number estimation illustrates the percentage that the tracking algorithm reports the number of the sources correctly. The cardinality error of the source number estimation ϵ_k is defined as

$$\epsilon_k = \sqrt{\sum_{j=1}^{J} \frac{1}{J} \left| \widehat{M}_{j,k} - M_k \right|^2}. \tag{42}$$

To evaluate the accuracy of source trajectory estimation, the position deviation under the correct number estimation is considered. Let $|\widehat{\mathcal{X}}_{j,k}| = |\mathcal{X}_k| = M_k$. The multiple speaker deviation at the jth MC run can be formulated as [41]

$$d_j(\widehat{\mathcal{X}}_{j,k}, \mathcal{X}_k) = \min_{\sigma} \sqrt{\frac{1}{M_k} \sum_{i=1}^{M_k} \left\| \widehat{\mathbf{x}}_{\sigma_i,k}^{j} - \mathbf{x}_{i,k} \right\|^2}, \tag{43}$$

where the minimum is taken over all permutations on σ. The mean deviation is given by

$$\xi_k = \mathbb{E} \left(d_j(\widehat{\mathcal{X}}_{j,k}, \mathcal{X}_k) \middle| |\widehat{\mathcal{X}}_k| = |\mathcal{X}_k| \right). \tag{44}$$

In general, equation (41) and (42) are able to give the performance of the source number estimation. Equation (44) gives the position estimation errors conditioning on the correct number estimation. By using these measures, the accuracy of both position and number estimation can be evaluated.

V. Experiments

In this section, both simulated reverberant environment and real room environment experiments are organized to evaluate the performance of the proposed algorithm. The parameters v and ρ in the source dynamic model are given in III-A. Since it is assumed that there is no prior information about the initial source position, this is

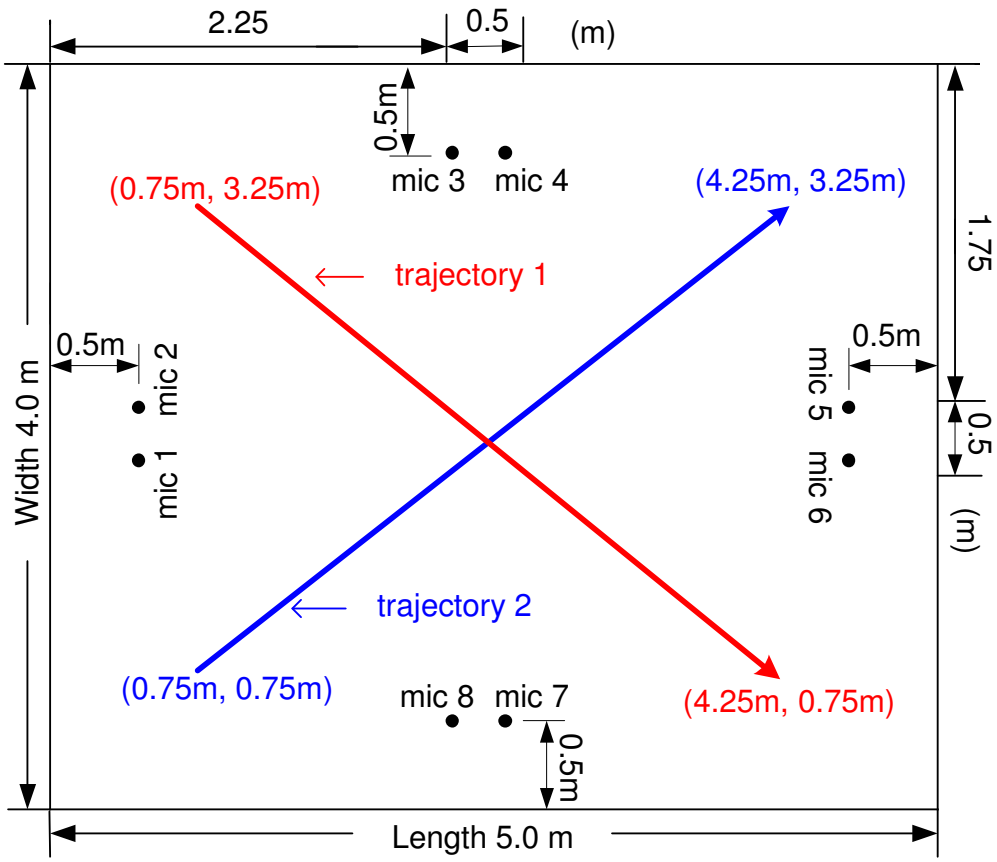


Fig. 3. Simulated room environment. Black dots numbered 1 to 8 denote the microphone positions, the solid lines represent the trajectories.

initialised at the center of the room with a velocity of 0.4m/s in both directions, i.e., $\mathbf{x}_0 = (2.5, 2.0, 0.4, 0.4)^T$. The corresponding initial variance is set as $\mathbf{P}_0 = \text{diag}([1, 1, 0.1, 0.1])$. The variance in the Langevin model is $\mathbf{\Sigma}_k = \text{diag}([1, 1])$. The measurement noise variance σ_τ is set to $5 \times (0.1)^9$, and 50 particles are employed for the proposed tracking approach. A small birth prior $P_\text{b} = 0.1$ is used to avoid an overestimation of the source number. The gamma parameter pair $(\alpha, \beta) = (4, 0.4)$ are found to be satisfactory in terminating the source track when the track is not associated with any of the TDOA measurements in a time step. These parameters are optimized by extensive experimental study and are found to be adequate in following experiments. The tracking performance for TDOA measurements using both DUET-GCC and traditional PHAT-GCC are demonstrated.

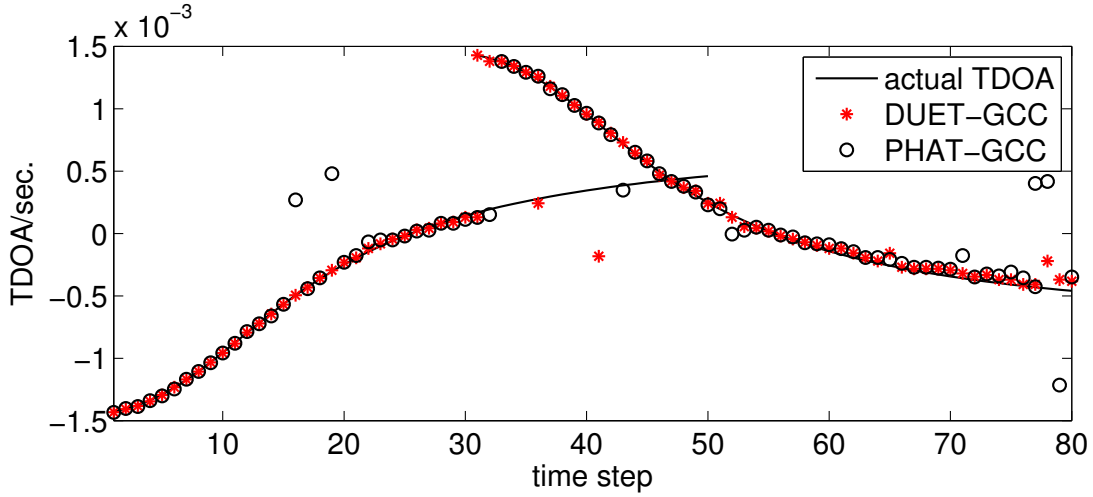
A. Tracking performance under a simulated room environment

Fig. 3 shows a simulated office room with dimensions $5 \times 4 \times 3\text{m}^3$. Four microphone pairs with a separation of 0.5m are organized around the center of the walls. The height of the microphones and sources are assumed to be known as 1.7m. The source motion trajectories follow two diagonal lines: one from bottom left to top right; the other from top left to bottom right. Two talkers appear at different times to form a time-varying number of

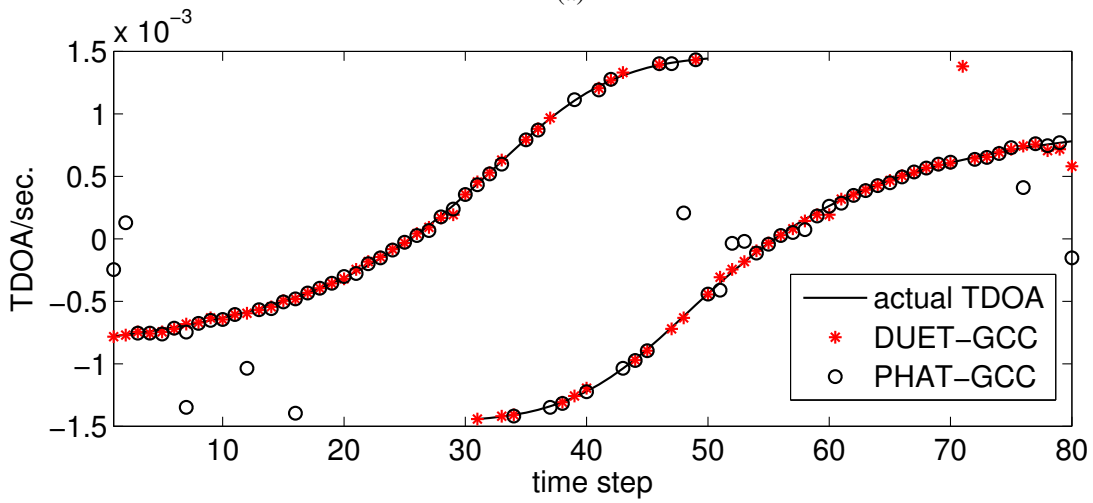
talkers: one is active from time step 1 to 50, and the other from 30 to 80. The speed of the source is set at 0.5m/s (1.8km/h), which is one third of a regular pedestrian walking speed, ranging from 5.32km/h to 5.43km/h [59]. Considering that a moving talker within a room is likely to be smooth and slow-paced, this experimental speed is reasonable and comparable with the source velocities in [10], [41]. The speech signals are processed with a frame length of 128ms, at a sampling frequency of 8kHz. Different wall reflection coefficients are set to simulate different reverberant environments. Noise conditions are simulated by adding additive white Gaussian noise (AWGN) with different variance. The room impulse response (RIR) is simulated by using the imaging method [60]. The received signal for a single source at each time step is obtained by convolving the clean speech signal with the RIR. These received signals are then added together to form the received signal for multiple sources. Three different tracking approaches are considered: 1) the proposed tracking approach using the TDOA measurements from the DUET-GCC method (DUET proposed); 2) the proposed tracking approach using the TDOAs from the PHAT-GCC method (GCC proposed); and 3) the RFS approach in [41] using the TDOA measurements from the DUET-GCC method (DUET RFS). In the implementation, the RFS tracking algorithm is exactly the same as that in [41]. The only difference is that the TDOA measurements are estimated from the DUET method. Hence, we name it as DUET RFS. The RFS approach based on the TDOA measurements from the PHAT-GCC method [41] is not implemented here as the PHAT-GCC TDOA measurements are much worse than the DUET-GCC based TDOA measurements. The GCC RFS in [41] will perform much worse than the DUET-RFS implemented in this paper. The number of particles for DUET RFS is set to be 200 and other parameters remain the same as used in [41].

1) Tracking results from a single experiment: In the first experiment, the results from a single experiment are presented. All the wall reflection coefficients are set to 0.6, which leads to a reverberation time $T_{60} = 0.163$s. The SNR is set to 30dB. To avoid exhaustive data associations, the TDOAs are extracted using the DUET-GCC method and PHAT-GCC method by setting the threshold values as 0.7 and 0.9 respectively to exclude false alarms. Figure 4 displays the TDOAs obtained from the microphone pair 1 and the microphone pair 2. The proposed DUET-GCC method presents better TDOA estimation for simultaneously active sources than the traditional PHAT-GCC method. The probability of detection can be improved and also, the false alarm rate can be reduced by using the DUET-GCC method. However, due to the reverberation and the interference between the source signals, it is very difficult to extract the TDOAs for multiple simultaneously active sources, particularly when two sources are closely spaced. For example, both methods degrade rapidly at microphone pair 2 when two sources are simultaneously active. Since the false alarm rate for the DUET-GCC and PHAT-GCC based TDOA measurements are different, the prior for false alarm is set to be different for two methods: $p_f = 0.05$ for the former one and $p_f = 0.1$ for the later.

Figure 5 shows the tracking result of a single trial from different tracking approaches. It shows that the proposed tracking algorithm with DUET-GCC based TDOA measurements is able to estimate the number of the active sources as well as the positions accurately. Although there are large measurement missing at the time steps when two sources are simultaneously active, the algorithm is still able to preserve the tracks, and thus lock onto the sources. The position tracking results are worse at the time steps when two sources are simultaneously active. The main reason for this degradation is that the TDOA measurements are not as accurate as those in the nonconcurrent



(a)



(b)

Fig. 4. TDOA estimates of (a) microphone pair 1; (b) microphone pair 2 from DUET-GCC and PHAT-GCC methods.

source scenario. The tracking loss likely happens when all or most of the microphone pairs fail to report correct TDOA measurements. In addition, false source detections are likely presented by using TDOA measurements from PHAT-GCC method. Given the same TDOA measurements, the RFS-PF approach [41] performs better at the time steps when only a single source exists in the tracking scene. This is because at these time steps, the TDOA measurements are estimated accurately. In addition, the RFS-PF algorithm under a single source scenario is similar to a PF algorithm which has been demonstrated to be robust in noisy and reverberant environment. However, when two sources are simultaneously active, the RFS-PF algorithm degrades significantly and performs worse than the proposed tracking approach. The reason is that serious miss detection happens at these time steps and the RFS-PF simply regarded the source as dead when none of the TDOA measurements is close to the ground truth. Hence, the

number of sources is underestimated and consequently, the tracking performance is deteriorated, while the proposed tracking method can still lock onto the sources by incorporating a more appropriate death prior.

To evaluate the performance over many MC runs, the measures defined in Section IV-C, the probability of correct number estimation P_k, the cardinality error ϵ_k, and the global mean deviation ξ_k are employed. Fig. 6 shows the average performance over 100 MC runs. The proposed tracking approach using the TDOA measurements based on the DUET-GCC method is able to provide good detection and tracking accuracy, even when the two sources are simultaneously active. The large error only presents at the time steps that the number of sources changes, i.e., the time steps when source birth or death happens. Again, the proposed tracking algorithm using DUET-GCC based TDOA measurement performs better than that using PHAT-GCC based TDOA measurements. For the PHAT-GCC method, it is very difficult for the tracking algorithm to detect the source due to large miss detection at some time steps. For example, at time steps 16, 31 and 71, the probability of correct number estimation is very small and all other errors are large. In addition, the performance of the proposed tracking algorithm is better than that of the RFS approach in [41]. It can be observed from Fig. 6(c) that the global mean deviation from DUET RFS is much larger than that from the DUET based proposed RFS method when multiple sources are simultaneously active. In general, the performance of the RFS approach [41] is favorably comparable with the proposed RFS approach when a single source is active, but performs worse when multiple sources are simultaneously active. The overall performance of the RFS approach is thus worse than that of the proposed RFS approach. It is worth mentioning that using the global mean deviation only can not fully illustrate the error of the position estimation since it is calculated based on results from correct cardinality estimation. In some cases, the algorithm can report accurate source position estimates, however, these position estimates are not counted into the global mean deviation due to an over-estimation or under-estimation of the source number.

2) Different simulated room environment: The algorithm is further implemented in a number of simulated experiments to fully study its performance. The experiments are organized based on different SNRs and different reverberation time T_{60}s. The simulated reverberation time T_{60}s are 0s, 0.163s and 0.289s respectively, and different noisy environments are 0dB, 10dB and 20dB. For the simulations under different T_{60}s, the SNR is fixed to 30dB; and for simulations under different SNRs, T_{60} is set to 0.1s. All the parameters of the tracking algorithm are set the same as those in the single trial experiment, except that under $T_{60} = 0.289$s and SNR=0dB, the priors of the false alarm are chosen as 0.5 and 0.55 for the DUET-GCC and PHAT-GCC based tracking respectively. Table II gives the average results over 100 MC runs. The tracking performance degrades as the noise or reverberation become heavier. This is mainly because the probabilities of detection in TDOA measurements decreases and the false alarm rate becomes larger. The tracking result based on the DUET-GCC TDOA measurements are better than that based on PHAT-GCC TDOAs under all experiments since the DUET-GCC method preserves the TDOA detections well and meanwhile excludes the false TDOA measurements more effective than the PHAT-GCC method. Further, the proposed tracking method with DUET-GCC TDOA measurements performs better than the DUET RFS method under different tracking scenarios.

B. Real recording experiment

The performance of the approaches is examined in a real laboratory located at the University of Edinburgh, Scotland. The room has carpet floor, concrete block walls and ceiling, one half-open wood door, and glass windows covered by hard cardboard with a thickness ≈ 0.4cm, as shown in Fig. 7. The measured reverberation time is $0.836s$ and the ambient noise level is -40dB [25]. The microphones are mounted on a set of T-bar stands, and the sources are set at a height of 1.33m. The microphone response is omni-directional within the frequency range 0 to 4kHz. The acoustic source used for all recordings is an omnidirectional speaker amounted on a small trolley, as shown in Fig. 7. The source is moved via a pulley mechanism and its position is measured using a laser measuring device, by which the sampled locations show that it is moving at a fairly constant velocity. The source signal is taken from the TIMIT database [61]. All measured signals are sampled at $f_s = 44.1$kHz and then downsampled to 8kHz, which is sufficient for acoustic source localisation and tracking (ASLT). The frame length is set to 1024 samples, or 128msec, and the source velocity is around 0.5m/s. The reverberant signals for two sources are recorded separately and then added together in the simultaneously active period to produce the received signals for a time-varying number of sources.

The measurements extracted from the microphone pair 4 (microphone 4 and 5) and the microphone pair 14 (microphone 17 and 18) are presented in Fig. 8. Due to heavy reverberation (the reverberation time T_{60} is as long as 0.8s as measured in [25]), the TDOA measurements are seriously deteriorated for both DUET-GCC and PHAT-GCC methods. The parameters in the tracking algorithm are set the same as in the simulated experiments except the prior of false alarms. Since the false alarms are heavier in the real audio lab experiments, the priors of false alarm are set as $p_f = 0.65$ and $p_f = 0.70$ for the DUET-GCC method and the PHAT-GCC method respectively. The tracking results from a single experiment is shown in Fig. 9. For the DUET-GCC measurement based tracking, the cardinality

TABLE II

TRACKING PERFORMANCE OF THE PROPOSED METHOD BASED ON THE DUET-GCC MEASUREMENT (DUET PROPOSED) AND THE PROPOSED METHOD BASED ON PHAT-GCC MEASUREMENTS (GCC PROPOSED) AND THE RFS METHOD [41] BASED ON THE DUET-GCC MEASUREMENT (DUET RFS) UNDER DIFFERENT ADVERSE ENVIRONMENTS.

	method	T_{60}			SNR		
		0s	0.163s	0.289s	0dB	10dB	20dB
P	DUET prop.	0.964	0.969	0.780	0.826	0.951	0.963
	GCC prop.	0.920	0.869	0.686	0.641	0.815	0.921
	DUET RFS	0.890	0.920	0.672	0.623	0.860	0.890
ϵ	DUET prop.	0.106	0.102	0.369	0.332	0.132	0.100
	GCC prop.	0.159	0.240	0.540	0.534	0.306	0.161
	DUET RFS	0.101	0.117	0.452	0.475	0.170	0.134
ξ	DUET prop.	0.086	0.103	0.381	0.256	0.109	0.093
	GCC prop.	0.100	0.185	0.494	0.439	0.136	0.112
	DUET prop.	0.241	0.234	0.481	0.459	0.248	0.221

estimation is much better than that based on PHAT-GCC measurement. This is because the TDOA measurements based on DUET-GCC method are more accurate than those based on the PHAT-GCC method, particularly when the two sources are simultaneously active, e.g., from time step 46 to time step 65. However, the tracking performance is worse than that in the simulated experiment due to a strong reverberation. It can also be observed that under the same TDOA measurements, the proposed tracking approach is able to provide better performance than the RFS method in [41].

To fully illustrate the average tracking performance for the real recording signals, the errors introduced in Section IV-C are presented in Fig. 10. The results show that the proposed tracking approach with DUET-GCC TDOA measurements is able to provide the best performance. The same as in the simulated environment, the performance is degraded at the time steps where source birth/death occurs. When multiple sources are nonconcurrently appearing, the tracking performance based on PHAT-GCC TDOA measurements is favorably comparable with that based on TDOA measurements from DUET-GCC method. The average results over 100 MC runs are given in table III. Under real lab environment, all these approaches are degraded significantly. However, the proposed tracking approach using DUET-GCC TDOA measurements performs the best in cardinality estimation as well as in position estimation.

VI. Conclusions

A TF masking based RFS-PF Method is developed to track an unknown and time-varying number of acoustic sources. Using the measurements extracted from DUET-GCC and PHAT-GCC methods, the performance of the tracking approach is fully investigated in the real audio lab as well as in the simulated room environment. The results from all experiments demonstrate that the proposed tracking approach is able to track multiple sources accurately. The experiment results also show that the tracking performance based on the DUET-GCC TDOA measurements are better than that based on the PHAT-GCC TDOA measurements. Also, the proposed tracking approach performs better than the RFS approach in [41].

However, tracking multiple acoustic sources in the room environment is still a challenge problem due to the reverberation and the interference among multiple source signals. For the tracking system developed in this paper, the number of acoustic sources is assumed to be small. An interesting direction for future work is to investigate the tracking approach for large number (say more than two) of sources. This unfortunately leads to following open questions. First, it requires more sophisticated approach to extract the TDOA measurements for multiple sources.

TABLE III

Tracking performance under real audio lab environment for three different approaches: DUET proposed, GCC proposed and DUET RFS.

error	P	ϵ	ξ
DUET prop.	0.765	0.390	0.304
GCC prop.	0.686	0.509	0.382
DUET RFS	0.651	0.496	0.379

This is not a trivial task since a short frame length is required to keep the system locking on the source dynamics, while extracting the TDOA measurements for multiple sources based on such short frames is very difficult. Further, assigning different hypotheses between the source states and the measurements will become computationally more expensive as the number of sources increases. One solution could be incorporating more advanced speech separation approach into our tracking framework to improve the estimation performance.

APPENDIX
EXTENDED KALMAN FILTERING

Following [22], the first-order Taylor expansion on the measurement function (37) is

$$\begin{aligned}\hat{\tau}_{n,k}(\ell) =& \hat{\tau}_{m,k-1}(\ell) \\ &+ \mathbf{C}_{m,k}(\ell)\left[\mathbf{x}_{m,k} - \mathbf{x}_{m,k-1}\right]^T + \bar{n}_k,\end{aligned} \tag{45}$$

where $\bar{n}_k = O_{\mathbf{x}}(\mathbf{x}_{m,k})$ is the higher order error of the time delay expansion, and $\mathbf{C}_{m,k}(\ell)$ is the coefficient vector of Taylor expansion

$$\mathbf{C}_{m,k}(\ell) = \frac{1}{c}\left[\frac{\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,1}}{\|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,1}\|} - \frac{\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,2}}{\|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,2}\|}\right]\Bigg|_{\mathbf{x}_{m,k}=\widehat{\mathbf{x}}_{m,k-1}}. \tag{46}$$

where $\widehat{\mathbf{x}}_{m,k-1}$ is the source position estimated at the previous time step $k-1$. Define

$$\bar{\tau}_{n,k}(\ell) = \hat{\tau}_{n,k}(\ell) - \hat{\tau}_{m,k-1}(\ell) + \mathbf{C}_{m,k}(\ell)\widehat{\mathbf{x}}_{m,k-1}, \tag{47}$$

where $\hat{\tau}_{n,k}(\ell)$ is the TDOA measurement extracted from the largest peak of the DUET-GCC function (10), and $\hat{\tau}_{m,k-1}(\ell)$ is calculated from (37). The nonlinear measurement is thus approximated by

$$\bar{\tau}_{n,k}(\ell) \approx \mathbf{C}_{m,k}(\ell)\mathbf{x}_{m,k} + \bar{n}_k. \tag{48}$$

Hence, the modified measurement $\bar{\tau}_{n,k}(\ell)$ is a linear function of the state $\mathbf{x}_{m,k}$ and a standard KF can be applied.

Assume that at the previous time step, the estimated state and variance are $\widehat{\mathbf{x}}_{m,k-1}$ and $\widehat{\mathbf{P}}_{m,k-1}$ respectively. Regarding (16) as the state process, the implementation of an EKF can be summarized as [62]

$$\mathbf{x}_{m,k|k-1} = \mathbf{A}\widehat{\mathbf{x}}_{m,k-1}; \tag{49a}$$

$$\mathbf{P}_{m,k|k-1} = \mathbf{A}\widehat{\mathbf{P}}_{m,k-1}\mathbf{A}^T + \mathbf{Q}\boldsymbol{\Sigma}_k\mathbf{Q}^T; \tag{49b}$$

$$\mathbf{S}_{m,k} = \sigma_\tau + \mathbf{C}_{m,k}(\ell)\mathbf{P}_{m,k|k-1}\mathbf{C}_{m,k}^T(\ell); \tag{49c}$$

$$\mathbf{K}_{m,k} = \mathbf{P}_{m,k|k-1}\mathbf{C}_{m,k}^T(\ell)(\mathbf{S}_{m,k})^{-1}; \tag{49d}$$

$$\widehat{\mathbf{x}}_{m,k} = \mathbf{x}_{m,k|k-1} + \mathbf{K}_{m,k}(\hat{\tau}_{n,k}(\ell) - \hat{\tau}_{m,k}(\ell)); \tag{49e}$$

$$\widehat{\mathbf{P}}_{m,k} = \mathbf{P}_{m,k|k-1} - \mathbf{K}_{m,k}\mathbf{C}_{m,k}(\ell)\mathbf{P}_{m,k|k-1}. \tag{49f}$$

After the EKF steps, the the posterior distribution is given by

$$p(\mathbf{x}_{m,k}|\mathbf{x}_{m,k-1}, \hat{\tau}_{n,k}(\ell)) = \mathcal{N}(\mathbf{x}_{m,k}; \widehat{\mathbf{x}}_{m,k}, \widehat{\mathbf{P}}_{m,k}). \tag{50}$$

REFERENCES

[1] M. Brandstein, "A framework for speech source localization using sensor arrays," PhD thesis, Brown University, Providence, U.S.A., 1995.

[2] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," PhD thesis, Brown University, Providence, U.S.A., 2000.

[3] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. 2000 IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 2, Jun. 5–9, 2000, pp. II909–II912.

[4] R. D. D. Zotkin and L. S. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *Proc. 2001 IEEE Workshop Detection and Recognition of Events in Video*, vol. 2, 2001, pp. 20–27.

[5] M. Brandstein and D. Ward, *Microphone Arrays. Signal Process. Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[6] F. Talantzis, A. Pnevmatikakis, and A. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 38, no. 3, pp. 799–807, 2008.

[7] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[8] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, Apr. 1997.

[9] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Process.*, vol. 2006, pp. 1–17, 2006.

[10] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[11] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sept. 2004.

[12] M. Fallon and S. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, May 2012.

[13] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.

[14] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[15] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Process.*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[16] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[17] U. Klee, Tobias, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP J. Applied Signal Process.*, vol. 2006, pp. 1–15, 2006.

[18] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Process.*, vol. 2006, pp. 1–19, 2006.

[19] E. A. Lehmann, "Particle filtering approach to adaptive time-delay estimation," in *Proc. 2006 IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 4, May 2006, pp. 1129–1132.

[20] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[21] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[22] X. Zhong and J. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2008, pp. 293–296.

[23] X. Zhong and J. R. Hopgood, "Particle filtering for time-delay of arrival based room acoustic source tracking: Mutiple nonconcurrent speakers," *Signal Process.*, vol. 96, pp. 382–394, 2014.

[24] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 5, May 2001, pp. 3021–3024.

[25] X. Zhong, "A Bayesian framework for multiple acoustic source tracking," Ph.D. dissertation, The University of Edinburgh, Scotland, U.K., 2010.

[26] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[27] S. Makino, T. W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, 2007.

[28] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, vol. 4, 2006, pp. IV–IV.

[29] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop Applications of Signal Process. to Audio and Acoust.,*, 21-24 Oct. 2007, pp. 275–278.

[30] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, Mar. 2008, pp. 4349–4352.

[31] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2914–2919, 1999.

[32] H. Christensen, N. Ma, S. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, April 2009, pp. 4593–4596.

[33] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb 2010.

[34] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1490–1503, Aug 2011.

[35] C. Liu, B. C. Wheeler, W. D. O'Brien Jr, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[36] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 4, pp. 806–815, 2013.

[37] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 520–529, Sept. 2004.

[38] I. Potamitis and G. Kokkinakis, "Speech separation of multiple moving speakers using multisensor multitarget techniques," *IEEE Trans. Syst., Man and Cybernetics, Part A*, vol. 37, no. 1, pp. 72–81, 2007.

[39] B.-N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.,*, vol. 2, May 2004, pp. 357–360.

[40] B.-N. Vo, W.-K. Ma, and S. Singh, "Localizing an unknown time-varying number of speakers: a bayesian random finite set approach," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 4, Mar. 18–23, 2005, pp. 1073–1076.

[41] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, 2006.

[42] N. T. Pham, W. Huang, and S. H. Ong, "Tracking multiple speakers using cphd filter," in *Proceedings of the 15th international conference on Multimedia*, no. 529–532, 2007.

[43] A. Masnadi-Shirazi and B. Rao, "An ICA-SCT-PHD filter appraoch for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 4, pp. 828–841, Apr. 2013.

[44] T. Otsuka, K. Ishiguro, H. Sawada, and H. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 493–504, Feb 2014.

[45] J. Taghia and A. Leijon, "Separation of unknown number of sources," *IEEE Signal Process. Letters*, vol. 21, no. 5, pp. 625–629, May 2014.

[46] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Process.*, vol. 16, no. 4, pp. 728–739, May 2008.

[47] M. Vihola, "Rao-blackwellised particle filtering in random set multitarget tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 2, pp. 689–705, April 2007.

[48] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *Proc. IEEE 15th Workshop on Statistical Signal Process.*, Aug. 2009, pp. 253–256.

[49] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2–15, January 2007.

[50] J. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 6, pp. 998–1003, Dec. 1982.

[51] J. M. Tribolet, "A new phase unwrapping algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, pp. 170–177, 1977.

[52] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged mutliple sensors," *Singal Process.*, vol. 87, pp. 1833–1847, 2007.

[53] D. Arnaud, F. N. de, and G. Neil, Eds., *Sequential Monte Carlo Methods in Practice*. Springer-Verlag New York, 2001.

[54] A. Doucet, N. de Freitas, K. P. Murphy, and S. J. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proc. 16th Conf. Uncertainty in Artificial Intelligence*, 2000, pp. 176–183.

[55] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004.

[56] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[57] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 4, pp. 1224–1245, Oct. 2005.

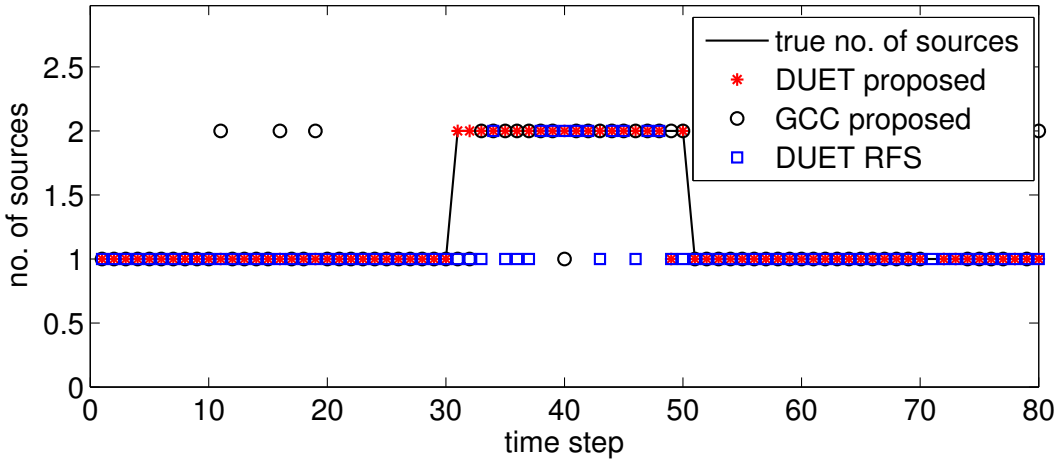
[58] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion.* Artech House, 2007.

[59] K. Aspelin, "Establishing pedestrian walking speeds," Portland State University. 2005. www.westernite.org.

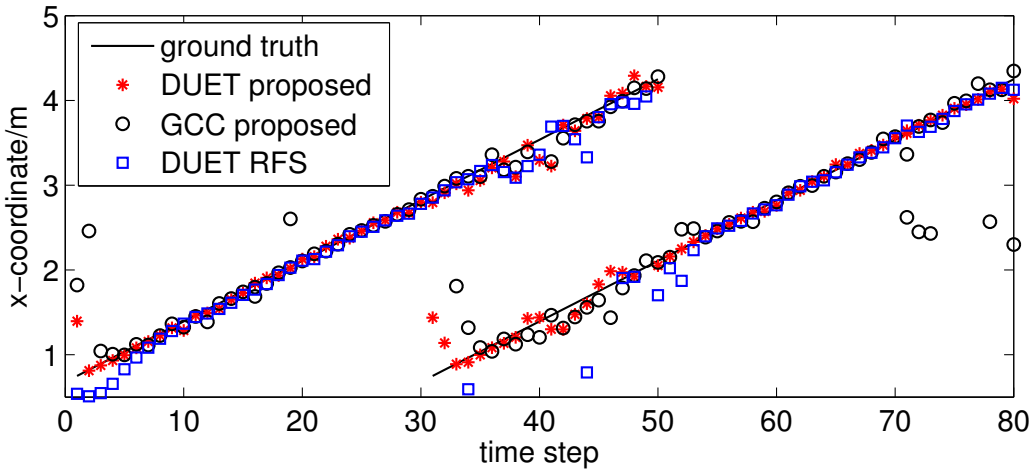
[60] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoust." *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.

[61] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993.

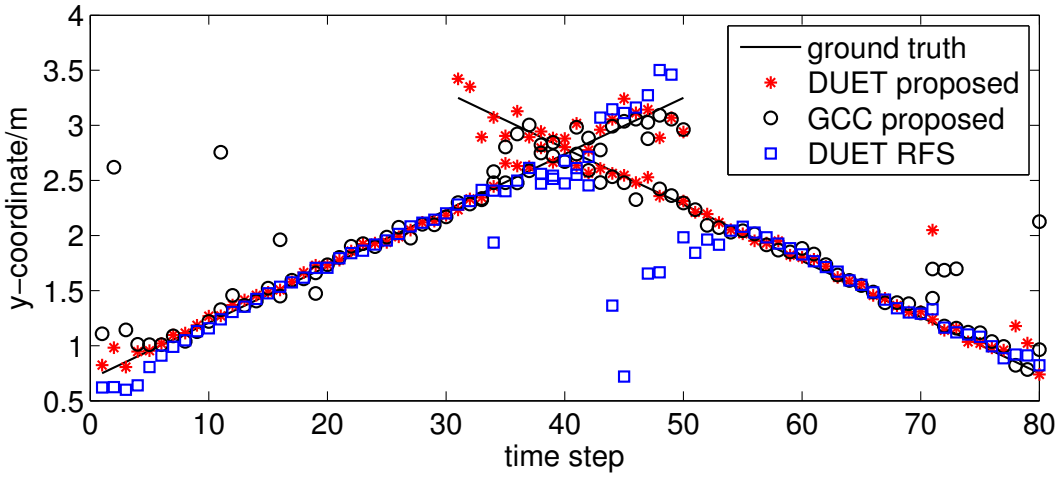
[62] D. Simon, *Optimal State Estimation*. John Wiley and Sons, 2006.



(a)

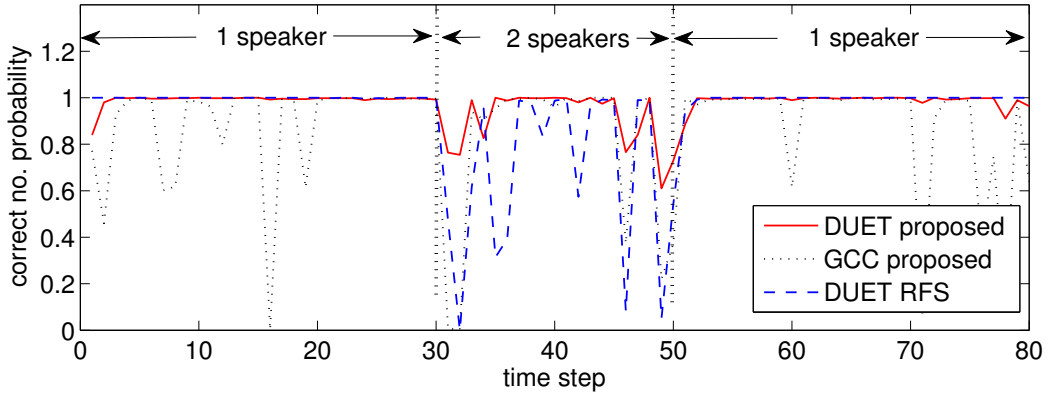


(b)

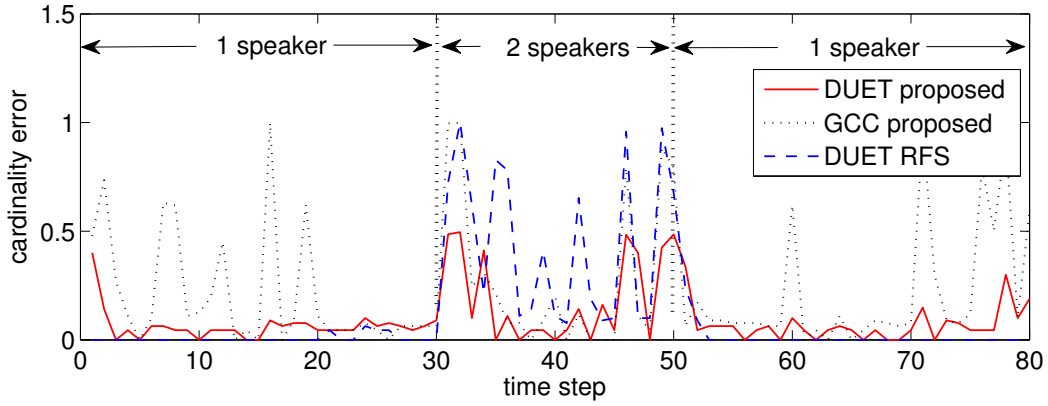


(c)

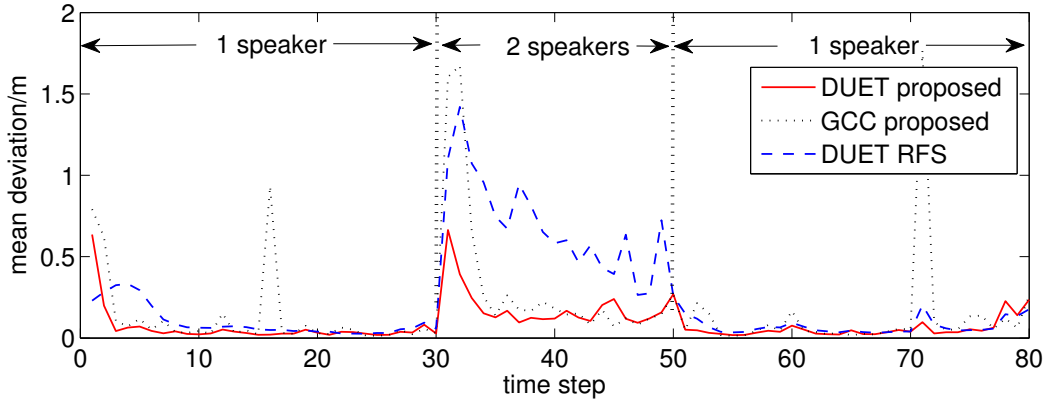
Fig. 5. Tracking result of a single trial under the reverberant environment ($T_{60} = 0.163$s). (a) Estimation of the number of the sources; (b) estimation results of the x-coordinate; and (c) estimation result of the y-coordinate.



(a)

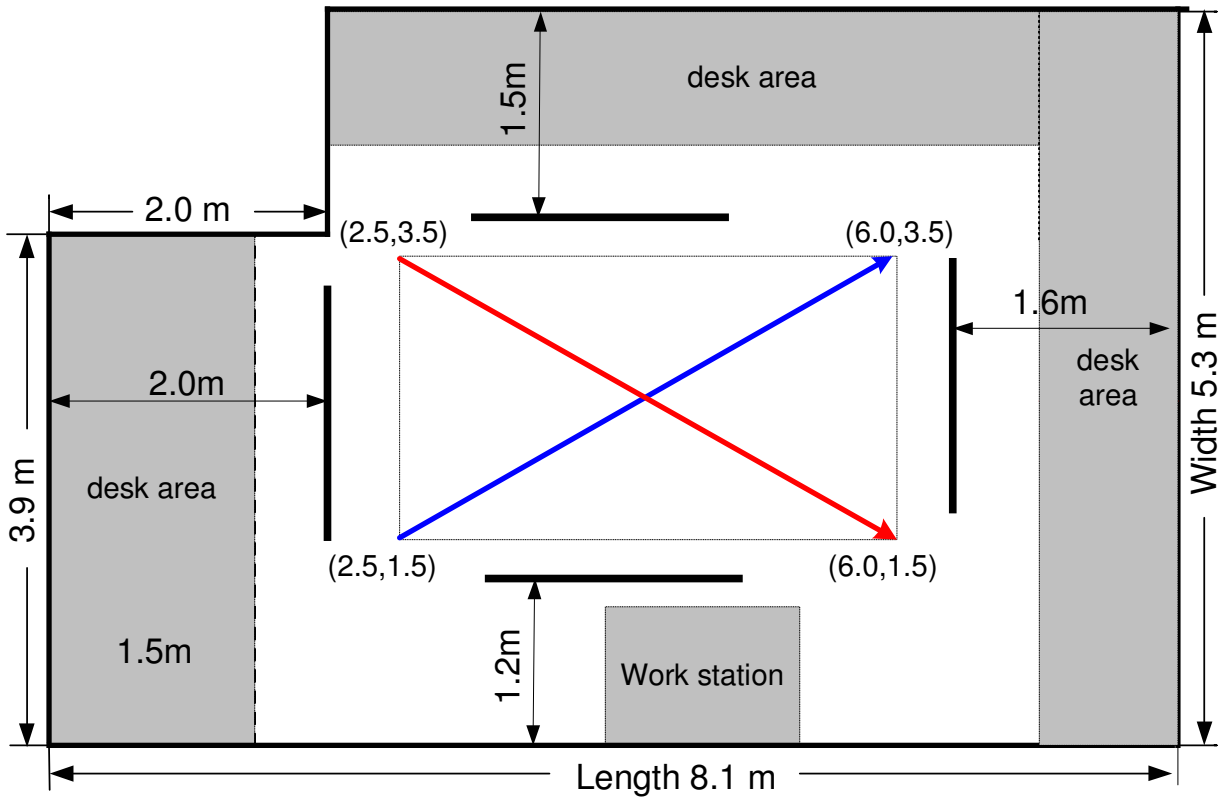


(b)



(c)

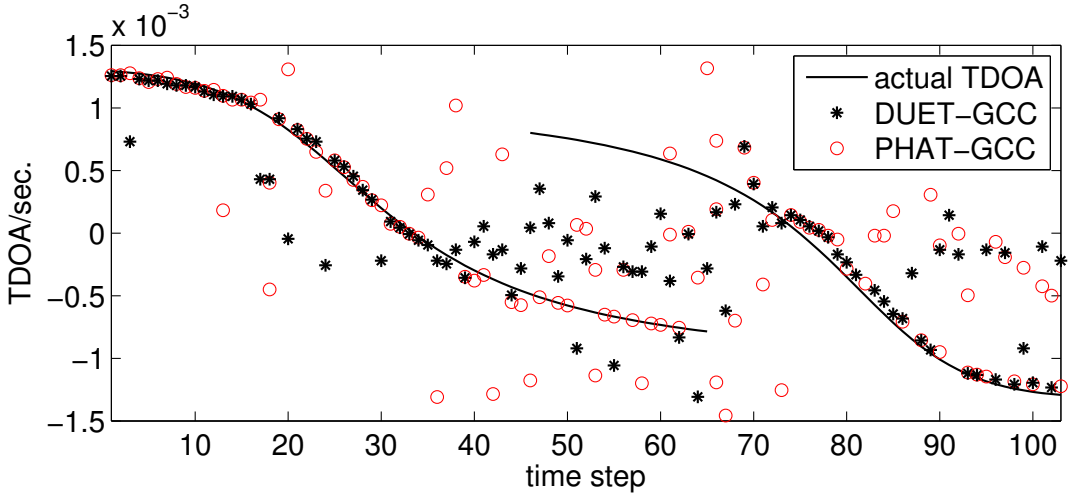
Fig. 6. Average tracking result of 100 Monte Carlo simulations under the reverberant environment ($T_{60} = 0.163$s). (a) Correct number estimation probability; (b) cardinality error; and (c) mean deviation.



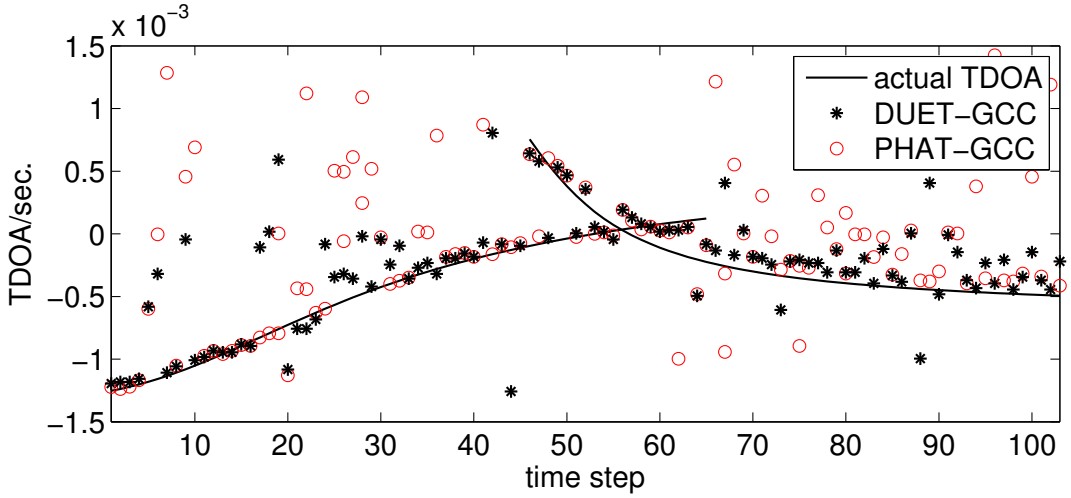
(a) Schematic: four microphone arrays shown in thick dark lines, each with five microphones. Sources move along diagonal lines.

(b) Real room environment and recording systems.

Fig. 7. Real audio room environment for performance evaluation.

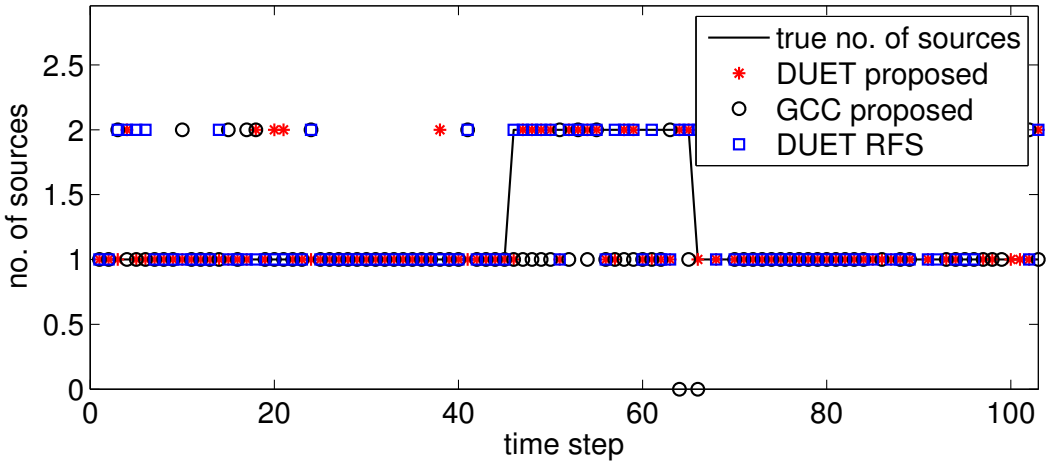


(a)

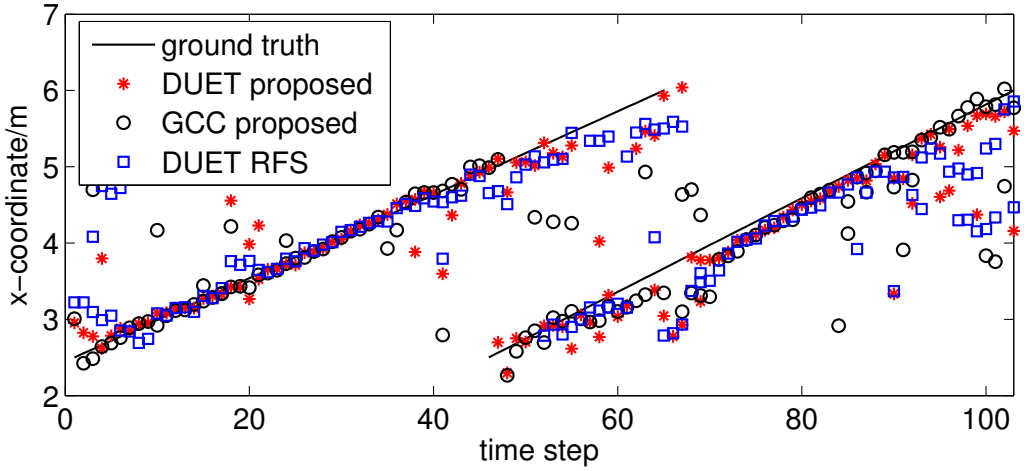


(b)

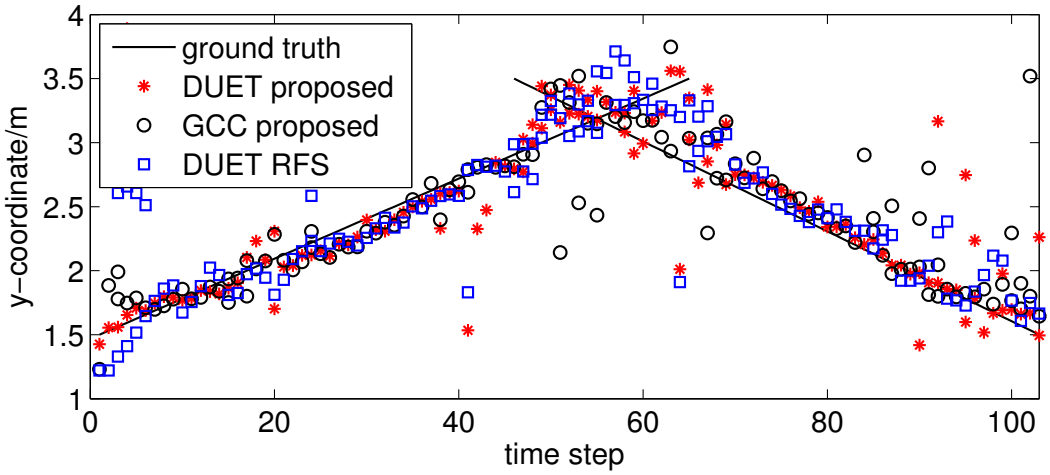
Fig. 8. TDOA estimates of (a) microphone pair 4; (b) microphone pair 14 from DUET-GCC and PHAT-GCC methods in the real audio lab environment. Source 1 is active from time step 1 to 65, and then source 2 follows from time step 46 to 103.



(a)

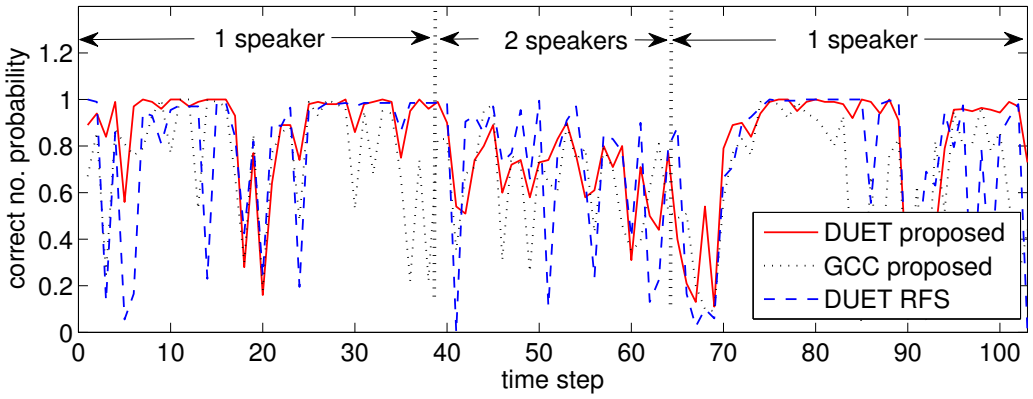


(b)

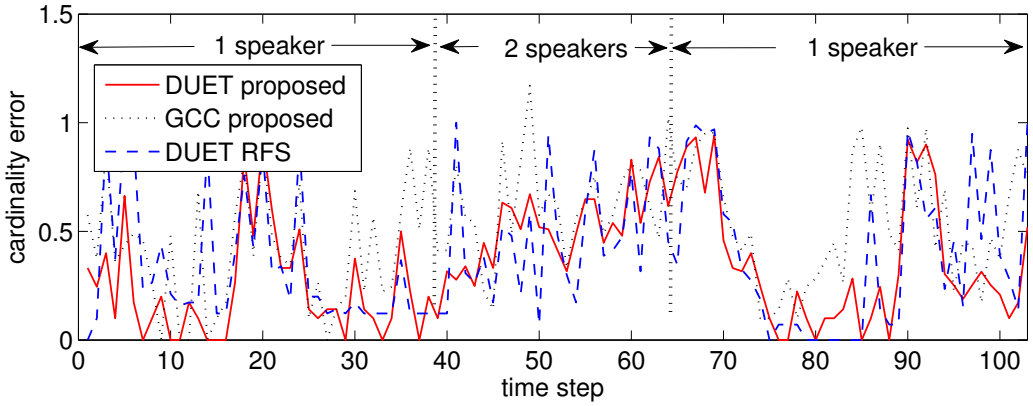


(c)

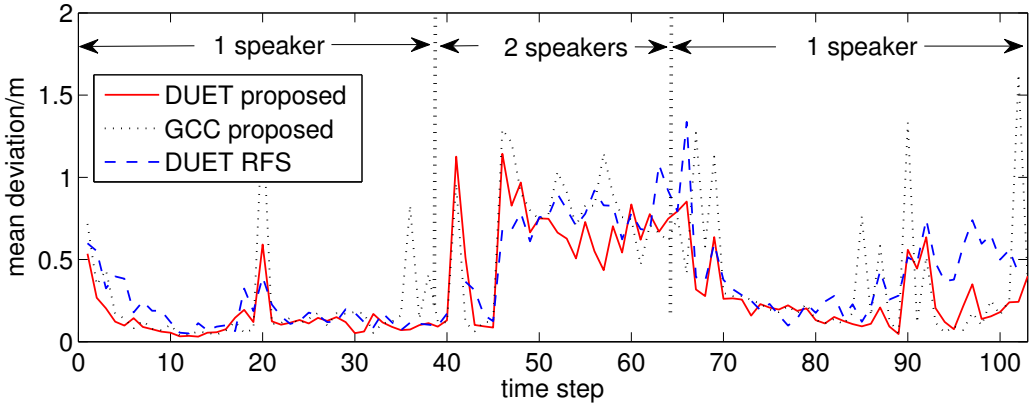
Fig. 9. Tracking result of the real recording signals. (a) Estimation of the number of the sources; (b) estimation results of the x-coordinate; and (c) estimation result of the y-coordinate.



(a)



(b)



(c)

Fig. 10. Average tracking result of 100 Monte Carlo implementations in the real audio lab environment. (a) Correct number estimation probability; (b) cardinality error; and (c) mean deviation.